資　料

Implementation and Assessment of Task-based Group Speaking-Tests for Large Low-Level EFL Conversation Classes

STURGEON Jason

1. Introduction

It is a well-known issue that English speaking tests are the most difficult type of test to organize and administer due to the strain on resources such as time, money, and personnel. A speaking test can only be useful if it does not make unreasonable demands on such resources, yet as a test, it must still maintain standards of validity and reliability (Bailey, 2005). This issue has been tackled in many ways by many instructors with varying degrees of success, ranging from the use of commercially available speaking tests such as the TOEIC speaking test, to interviewing each student individually with a predetermined list of questions.

The issue of reliability is particularly difficult since speech is often imprecise by nature. Even native speech is filled with stops, restarts and small errors, and it's not uncommon for someone to stutter or change their mind mid-sentence. The challenge that speech evaluators face is how to quickly and fairly decide whether such mistakes that even native speakers make should count against a student's score or not. Furthermore, if multiple raters were to grade the same test, the end results are likely to vary (unless all raters have been highly trained), leading to low reliability.

Limitations on test time are also a large problem. If all students are to be tested within a given amount of time, as is the case in most institutions, the larger the class, the shorter each student's test must become. For example, if each student in a class of 30 were to undergo a speech test with the teacher in a single 90-minute class period, each student would only have 3 minutes to speak. This, combined with the additional time required between tests (students will likely need to switch seats, receive instructions from the teacher, etc. before the test begins), 3 minutes becomes an absurdly short amount of time, and meaningful assessment becomes impossible.

Possible solutions include using multiple proctors, decreasing the number of students per class, lengthening the time allotted for the test, testing the students in groups instead of individually, and making use of computer software. Each of these potential solutions has a drawback, however. Involving multiple test proctors, decreasing students, and lengthening test time are generally outside of the control of the teacher where universities are concerned, which removes those as options. Group tests can solve the issue of time but add a new issue with test validity and fairness, since students of varying levels of skill would use the test time unevenly to speak; one student may dominate the conversation while another may not get a chance to speak at all. Computer based testing generally involves speaking prompts either viewed on a screen or heard through speakers, followed by the testee's response, which is recorded. This also solves the

連絡先　STURGEON Jason　〒401-0380　山梨県南都留郡富士河口湖町小立7187　健康科学大学　健康科学部　作業療法学科　Tel：0555-83-5200　E-mail：jsturgeon@kenkoudai.ac.jp

time problem, but the recordings must be judged by a human, which means reliability is still an issue.

To tackle these issues, I made use of task-based learning assessment (TBLA), which takes a different approach to speech assessment. To address the issue of reliability, I discarded the traditional idea of assessing a student's speaking proficiency via subjective criteria such as pronunciation, complexity, lexical density, and accuracy, and adopted objective criteria based on the successful (or unsuccessful) completion of tasks. A student receives points based on how well they meet the requirements of a task in a set amount of time. Tasks can be designed with very clear outcomes, making their scoring much more straightforward and reliable, and the speed at which they are able to accomplish these tasks serves as an apt measure of communicative competence.

While the concept of TBLA is not new, it remains largely unused. Shehadeh (2018) explains, "many teachers do not know how to utilize TBLA in their practices" and that "[they don't]...know what TBLA is exactly and why it is more conducive to L2 learning." According to Ellis (2003), "Task-based testing is seen as a way of achieving a close correlation between the test performance, i.e, what the testee does during the test, and the criterion of performance, i.e, what the testee has to do in the real world." This implies that it better reflects the actual function of speech used in authentic situations when compared with fluency-based testing.

While the various functions of language have been defined many times by many scholars (see Finch 1997, Leech 1974, Halliday 1975), one overarching theme emerges; we use language to effect and influence the world around us and the people in it. It is reasonable to assume that if someone achieves a desired result through language (by getting the information they wanted or the thing they needed), then they have successfully used that language, even if it contained some errors. This is especially true for both children and beginner-level students who have little speaking experience.

At these lower levels of skill, I argue that learning and achievement should be measured by how quickly and effectively the student is able to navigate language scenarios, as this is closer to how the real world operates and is therefore more practical to the student. Real conversations, as opposed to "classroom conversations", do not have the luxury of time to ponder over the precise wording or grammar of a sentence; people will forgive small lapses in pronunciation and grammar as long as they can get the gist of the messages, but they are less forgiving of awkward pauses that drag on and impede the flow of conversation.

For the issue of time constraints, the tasks I have created are designed to be completed by groups of 3 or 4 students at once, allowing for up to 4 times as many students to be tested than if they were tested individually. Issues of fairness and validity are taken into account via the test design.

2. The Speaking Test

2.1 Test Set-up and Flow

A single group's test takes about 30 minutes. The speaking test is held in groups of 3 or 4 students around a table with a 4-way partition in the center, tall enough to prevent students from seeing each other's test papers, but short enough so that they can see each other's faces. Each seat has clear label, 1~4, to help facilitate speaking-turn order. A sensitive microphone (ideally a condenser microphone with an omnidirectional setting) is placed in the very center of the table (above where the partition intersects) equidistant from each student and is used to record the entire test. A video camera is placed off to the side, at an angle that keeps all students in frame, and likewise records the entire test.

Once the students are seated (in whichever number seat they like), the test proceedings are described to confirm the students' understanding (5 minutes). These include the following:

- The test is recorded
- Their seat number dictates their speaking turn order
- The time limits of the test
- The penalties for using Japanese during the test
- Time allowances for answering clarification questions from other students
- The allowance and encouragement of other communication strategies, such as gestures

Test materials, including answer sheets and information-gap task prompts are distributed to each student based on their seat number (note that students do not know the content of their test materials when they choose their seat), and the goals of the task, as well as how points are earned and what is required to pass the exam are explained. Both the audio and video recordings are started, and the first task begins with the student in seat 1.

The instructor uses a timer with a pause function to track speaking turns. The length of time each student receives can be varied as needed, but tests involving 3 tasks with each student receiving a 2-minute turn per task, have thus far been successful.

The instructor begins the timer, and student 1 conveys the information on their test paper to the group as quickly and effectively as they can, while the other students listen and try to fill in the missing information on their own test papers based on what they hear.

If a student has not understood something, they may ask student 1 questions, at which point the instructor pauses the timer to allow student 1 to respond. Once the response is given, the timer is resumed. In this way, questions from other group members will not impede or affect student 1's own score. This method also reflects people's ability in real conversation to stop someone and ask them questions.

Once the timer runs out, the instructor abruptly stops student 1, and moves on to student 2, giving them 2 minutes to complete their task. This repeats for all remaining group members.

Once student 4 completes their 2-minute turn, the instructor distributes the materials for the next task (if they have not done so already), explains it, and carries out the next task in the same way. "Free flowing" tasks have sometimes been implemented, in which there is a single timer applied to the group as a whole, with the length of the timer based on the number of group members (e.g, if number of member×2:00 is used, a 3-student group would receive 6 minutes, a 4-student group would receive 8 minutes, etc.). In these tasks, each student has the same amount of information to share, however they may speak whenever they like without the constraint of waiting for their turn.

Once all tasks have been completed, the instructor collects the test papers, stops the recordings, and saves them to be graded later, and ushers in the next group to be tested.

2.2 Test Content

The content of the test is based off of TBLA activities and conversations that the students have learned about and practiced in class. It uses a wide variety of information-gap activities in which each student has a few pieces of information that the other students do not have, but when all 4 students' information is combined, it completes the whole picture or story, sometimes referred to as a "jigsaw" activity. Students, using only English, must ask each other questions about incomplete areas on their test sheets, and fill in the gaps accordingly.

An example of this is a map navigation task. In such a task, all 4 students have the same map of a city,

however on each test, several buildings are not labeled. Student 1 may be able to see the location of “Statton High School”, but the school is missing or is not labeled on student 2, 3 or 4’s sheets. Student 1’s task is to describe the location of the school, or perhaps a route to the school from some other common-point to the other 3 students, while those 3 students must listen and comprehend student 1’s explanation well enough to find that location on their own map and label it accordingly. On any given test, there will be more areas of missing information that can possibly be covered in 2 minutes, but this is intentional, and allows students to demonstrate how quickly and efficiently they can communicate.

Other themes include describing people or objects, comparing schedules to make plans with friends, asking about menu items, find-the-difference tasks, and more. Information-gap tasks are useful because they can be adapted to nearly any theme, including English for Specific Purposes.

3. Evaluation

Students are evaluated on both speaking efficiency and listening skill. Tests are scored and evaluated by comparing the student’s recorded audio/video data and their test sheets. The evaluator (a native or high-advanced level speaker) uses a blank “base version” of the task (one which does not have any of the 4 students’ information on it) to assess a student’s speaking ability. It is important to use a blank test to avoid evaluation bias, as the evaluator could easily be influenced if they can see what the student might try to say before they say it.

3.1 Speaking

The evaluator starts by listening to the recording of a student’s speaking turn from start to finish, following all directions and descriptions as the student says them in real time, and marks the blank sheet accordingly. Afterward, they compare the result they got through listening to the student’s actual test sheet and compare the accuracy of what they heard the student say to the information the student was supposed to say. For each test item that was correctly communicated, the student receives a point, and that becomes a mandatory listening question for the other students in the same group (i.e., the item was sufficiently described or explained by the student, so the other students should have the same answer as the teacher if their listening skills were proficient enough). If the student’s description led the teacher to an incorrect conclusion, caused confusion, was incomplete, or left room for doubt, it was not counted towards their score. In these cases, whatever the other group members may have written on their test in response to this invalid description cannot be expected to be correct, and so these responses are ignored, and incur no penalty.

Each task has a predetermined minimum item threshold (determined by the instructor in advance) that must be reached within the time allotted in order to achieve a perfect score. For example, in the map navigation task mentioned above, students are required to describe the location of 3 buildings that they could see within their allotted 2 minutes to receive full marks. If students are able, they may continue to describe items above and beyond the minimum number for bonus points to compensate for other errors they may have made (described below).

3.2 Listening

For listening, scoring is a subtractive process. Students begin will full marks for listening, but may lose points as the test progresses if they are unable to fill in their test sheets accurately according to their classmates’ instructions.

As described above, students are expected to understand the directions and explanations of other students whenever the evaluator is also able to do so. As such, whenever a student misrepresents information on their sheet despite it having been

sufficiently explained, that student will lose points for listening. Students are encouraged to take responsibility for their own understanding through the inclusion of the question-response time allotment mentioned above. Students are clealy informed that if they are not sure what was said, it is their responsibility to talk to the speaker to obtain clarification. The fact that the timer will be paused in such cases is again mentioned here to mitigate the potential worry of haveing one's turn interrupted.

3.3 Further Score Adjustments

After an initial score is calculated, other bonuses or penalties may be incurred, based on the goals of the test. Penalties are imposed for each use of a Japanese utterance that carries meaning relevant to the test, and to a lesser extent, penalties for Japanese-English "buzzwords" that are often mistaken for English, but are not, such as using the word "smart" to mean "thin and well dressed". That said, students are not penalized for saying "uhh" or other irrelevant self-talk in Japanese.

Finally, if the student was able to communicate in such a quick and efficient manner that they covered more than the minimum number of test items, each of those bonus items is allowed to cover a penalty they incurred from listening or Japanese use (if any). Other teacher might implement other types of bonuses or penalties based on the goals of the speaking class, in hopes of positive washback.

4. Feedback from Students

Regardless of the researcher's opinions of this group testing method, it is unlikely to be useful to other teachers unless students themselves also feel that the test is fairly scored, relevant to their lives, and provides an accurate measurement of their speaking skill. If students do not like the test, it will only be a source of anxiety and negative washback.

4.1 Student Questionnaire

While the feedback received from the course evaluation questionnaire each semester has consistently been overwhelmingly positive (scoring 4.5 or higher out of 5), the questionnaire provided by the university only asks questions about the course as a whole, and does not ascertain student opinions regarding how the speaking test is handled. Therefore, a separate evaluation was needed. An online questionnaire asking 6 questions (a combination of Likert scale and multiple choice) about specific test aspects, and a space for an optional open response prompt was created to obtain students' opinions about the test. These questions (translated in English) were as follows:

1. What year of school are you in?
2. How do you feel about the test emphasizing actual conversation over the correct use of grammar?
3. How difficult did you find the test?
4. How fair did you find the final test (in groups of 3 or 4) to be?
5. Do you think the English Conversation Test was an accurate assessment of your actual English communication skills?
6. What do you think about allowing alternate communication strategies (gestures, etc.) in the English conversation exam?
7. Other thoughts or opinions (open answer).

A message was sent via Microsoft Teams to all students currently enrolled at the university who had taken the English conversation course in the past (n=147), asking for volunteers to take a short anonymous questionnaire regarding the English conversation test. This message informed students that participation was voluntary and held no bearing on their grades or other assessments. Those students who wished to participate could click on a link to the online questionnaire. The informed consent page was displayed at the start of the questionnaire which further described the study, and again informed the student that participation was

voluntary and completely anonymous, their names and other personal information would not be known even to the researcher. After reading the informed consent agreement, students were given the option to either participate by filling out the questionnaire or not participate by simply exiting. All informed consent related communications were in the students' native language.

4.2 Questionnaire Results

The questionnaire had a response rate of 49% (n=72 out of 147).

For item 2, 1.4% of students "disliked" the idea of being graded based on conversational ability rather than correct grammar usage, 9.7% had "no opinion", 41.7% "somewhat liked" it, and 47.2% "liked" it.

On item 3, 2.8% of students said the test was "very difficult", 13.9% said it was "somewhat difficult", "70.8%" said it was "just right", 8.3% said it was "somewhat too easy", 1.4% said it was "way too easy", and 2.8% "didn't remember".

When asked about the fairness of the test, 1.4% found the test to be "not fair at all", 31.9% thought the test was "somewhat fair", 36.1% answered "mostly fair", 22.2% said "totally fair", and 8.3% could not remember.

When asked if they thought the test was an accurate measure of their real-world speaking ability, 5.6% responded "somewhat disagree", 44.4% were "unsure", 34.7% "somewhat agreed", and 15.3% "agreed".

For item 6, 1.4% of students thought communication strategies should be "prohibited", 11.1% though they "should be limited", and 87.5% felt that they should be "allowed".

Finally, there were 10 students who provided a response to the open opinion question, however, 8 of the responses were not particularly useful to this research, and included short statements such as, "thank you", "no comment", or "it was fun". Of the other two comments, one student responded that they did not feel that communication strategies and word approximation would work in a conversation with a real foreigner. The other student commented that they would like to see a patient/doctor scenario included in the test to make the test more relevant and useful for the students' future careers.

5. Discussion

Here I will attempt to answer some of the questions raised above using the data collected from the questionnaire in conjunction with my own experiences.

5.1 Reliability

After conducting the task-based test described above more than 20 times over the last 6 years, I am convinced that it is a reliable method. When listening to student recordings, as long as the evaluator is not aware of what the student's communicative "target" is, he or she can provide a non-biased interpretation of what they hear, following what the student says exactly. This results in a well defined "answer" on the evaluator's test sheet (e.g, a location on a map, a character in a line-up, a drawing of a complex object, etc.), which can then easily be compared to the student's sheet afterward. If the target and answer match, the task was successfully completed. If the answer was unclear, or diverged from the target, they failed the task.

While the language students used to accomplish tasks varied a great deal in complexity, fluency and accuracy, the result was always binary; they either spoke in a manner that was coherent enough for me to understand, or they didn't. They spoke with enough accuracy to help me arrive at the correct result, or they didn't. I found that no matter how many times I re-graded a student's test, it never fluctuated, which provides evidence that this method of testing is very reliable.

5.2 Time Resources

As mentioned above, time is a big concern for any

speaking test. Currently, I use tests which include 3 8-minute group tasks, and schedule groups of 4 students into 30-minute slots. This allows for 12 students to be tested in one 90-minute period (or 24 in 2 periods, which I often do). While this is not the fastest method of testing, it is a bit faster than individual testing, and offers the option of adding or removing tasks from the test to make it longer or shorter as needed. Increasing or decreasing the group size also provides some measure of time control, as having fewer groups leads to less time spent explaining test procedures and moving in and out of classrooms when groups switch.

While changes to the test can be made to affect the amount of time required, it must be pointed out that the amount of time needed for grading will be directly proportional to the amount of audio/video data recorded during the test, as the teacher must listen to the recording from start to finish, in essence taking the test themselves as a listener. In many cases, grading will take slightly longer than the actual test, as the teacher may need to adjust the volume levels on their playback device and listen more than once. However, the teacher may grade the tests at their leisure since the data may be played back anytime. While slow, this aspect of the grading process allows for even greater reliability and accuracy.

5.3 Validity

As we have seen, task-based tests are one of the closest means of testing we have to actual real-world language use. Item 5 of the questionnaire lends some evidence to this claim, as 50% of students thought that the test was an accurate or mostly accurate measure of how they would perform in a real-world situation. Another 44.4% were unsure. This large number may be because they lack real-world experiences using English and have no basis for comparison. More importantly, only a small percentage of students (5.6%) felt that the test and the real-world were unrelated.

5.4 Student Receptiveness

Results of the survey indicate that students are generally satisfied with this testing method. Responses to item 2 in which they were asked how they felt about a conversation-focused rather than grammar/accuracy focused test were positive, at 88.9%. This data alone gives ample evidence to suggest that students are more than willing to be tested in this way, which may lead to positive wash-back in the classroom.

Item 6 regarding the allowance or prohibition of communication strategies to help them communicate during the test was also strongly in favor of their inclusion (87.5% in favor, with another 11.1% in favor with some limitations), indicating that low-proficiency students feel that such strategies are useful and important for communicating effectively.

70.8% of students felt that the difficulty of the test was balanced, which is what we would hope for statistically when considering how difficult to make a test. For this question, student responses followed normal distribution closely.

Finally, 58.3% of students responded positively when asked about how fairly they felt the test was handled, with another 31.9% responding “somewhat fair”. There is an inherent issue with using pairwork or groupwork as an assessment method, which is that there is a chance that poor performance on behalf of one student will negatively impact the performance or scores of the other group members. If not handled fairly and appropriately, this issue could cause a great amount of frustration for students who are genuinely skilled, but happen to be paired with someone who isn’t, leading to demotivation. While the majority of students seem to find the test fairly scored, the 31.9% who only responded “somewhat fair” is concerning. This may indicate that while they feel the test is generally fair, there are certain aspects of it that could be handled better, resulting in mixed feelings. This result may warrant further investigation into what aspects of the test students found unfair, and what might be done about it.

5.5 Limitations

The participants in the study were low-level English speakers, so it is unclear if the results of this study will generalize very well to higher level students, who may feel differently about this testing method. Also, 44.4% of students were unsure if the test was valid, which is a substantial percentage. Depending on the reasons for this uncertainty, the interpretation of the results might significantly shift. The data alone however is not enough to draw any conclusions from this figure.

6. Conclusion

The task-based group speaking test discussed in this paper provides a number of benefits, but has a few shortcomings. Benefits of using this method include a high-degree reliability, and according to research, validity in terms of applicability to the real-world. Most of all, it is highly rated and accepted by students for being communication-focused rather than accuracy focused, for being well balanced in difficulty, and for allowing communication strategies which would be available to them in the real world. Its main drawback is the time required to administer and grade the test, which may or may not be feasible for some teachers depending on their workload. There also may be some concern with the overall fairness of the test according to student data.

Suggestions for improving this testing method include investigating the possible existence of unfair test procedures and correcting them, incorporating course content relevant to students' majors that will feed into test content, and looking for alternative time-management schemes.

References

Bailey, K. (2005). *Practical English language teaching: Speaking*. New York, NY: McGraw-Hill.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford, England: Oxford University Press.

Finch, G. (1997). *How to study linguistics: A guide to understanding language*. New York, NY: New York University Press.

Halliday, M.A.K. (1975). Learning how to mean. In E.H, Lenneberg & E. Lenneberg (ed.). *Foundations of language development* (pp. 239-265). London, England: Academic Press.

Leech, G. (1974). *Semantics*. England: Penguin Books Ltd.

Shehadeh, A. (2018). Task-based language assessment. In J.I, Liontas (ed.). *The TESOL encyclopedia of English language teaching* (pp. 1–6). Hoboken, NJ: Wiley-Blackwell.