International Journal of English Language Teaching

Vol.5, No.3, pp.27-40, April 2017

Published by European Centre for Research Training and Development UK (www.eajournals.org)

ENGLISH SPEAKING ASSESSMENT: DEVELOPING A SPEAKING TEST FOR STUDENTS IN A PREPARATORY SCHOOL

Zhanna Shatrova[1*], Rasheedah Mullings[2], Stěpánka Blažejová[3], Eda Üstünel[4]

1. *School of Foreign Languages, Mugla Sitki Kocman Univerisity, Turkey, Mugla 48000, Turkey*
*e-mail: zhannashatrova@mu.edu.tr tel: (90) 545-902-3405
2. *School of Foreign Languages, Mugla Sitki Kocman Univerisity, Turkey, Mugla 48000, Turkey*
e-mail:rasheedahmullings@gmail.com
3. *School of Foreign Languages, Mugla Sitki Kocman Univerisity, Turkey, Mugla 48000, Turkey*
e-mail: stepankablazejova@mu.edu.tr
4. *School of Foreign Languages, Mugla Sitki Kocman Univerisity, Turkey, Mugla 48000, Turkey*
e-mail: eustunel@mu.edu.tr

ABSTRACT: *This article discusses the format for the speaking exam developed for the School of Foreign Languages at a Turkish university. The introduction and literature review provide the rationale and arguments for the suggested changes to an existing format. The areas for review and redesigning are: the structure of the exam, assessment rubric, exam questions, and supporting manual for assessors and interlocutors. Each area is considered from the standpoint of best practices in speaking skills assessment and the ways those practices should be utilized in a particular learning community. A detailed description of the procedures, logistics, and sample questions follows.*

KEYWORDS: testing speaking skills, assessment, interlocutor, speaking task specification

INTRODUCTION

Speaking is a complex matter. To be considered fluent the learner should reach the level of automatic formulation and production of statements (Fulcher, 2003). With advances in the knowledge economy, the growing demand for specialists proficient in English, spread of English as a Medium of Instruction and English for Specific Purposes practices at universities worldwide, the issue of speaking skills assessment comes to the forefront. ESL programs and schools in colleges face the challenge of designing speaking tests which best meet the needs of their learners and department requirements.

Assessment of learners' speaking proficiency has always been "in a state of flux" (Norris 1997, p. 148; Fulcher, 2003). The majority of standard tests such as Test of English as a Foreign Language (TOEFL) and the Cambridge ESOL (English for Speakers of Other Languages) Examinations, including the International English Language Testing System (IELTS), assess four language skills. Speaking takes 25 percent of the total grade; so speaking skills are not specifically the focus of the evaluation. The tests that are designed to test oral proficiency, such as the Trinity College London Graded Examinations in Spoken English (GESE) and the Test of Spoken English (TSE), aim at assessment of test takers' ability for oral communication for employment and certification purposes. Roca-Valera and Palacious (2013) pointed out the wash backs of these tests: not very clear which language features are under consideration, no mechanisms to adapt to teaching methodologies and specific preparatory courses in colleges and universities; more focus that should be given to test takers and test assessors (p.66).

The American Council on the Teaching of Foreign Languages (ACTFL) (1986) and Common European Framework of Reference for Languages (CEFRL) (2011) are seminal guidelines which address all aspects of speaking skills assessment and levels providing detailed description and explanation of language learners' competencies, common reference levels, and categories for the description of language use. Semi direct testing and Oral Proficiency Interview (OPI) have been viewed as two main types of speaking tests. A number of studies on these types in the 90s addressed the issues of validity and reliability of these test approaches in assessment of speaking skills in different languages. In a study on a Portuguese speaking exam which included both semi testing and OPI, Stansfield, Kenyon, Doyle, Paiva, Ulsh, and Cowles (1990) found evidence that though the validity of both tests was almost equal, the results were more consistent with semi direct testing. Taping had advantages when it came to time constraints. It allowed for testing more students in a shorter time. At the same time, examinees preferred OPI because of a greater level of comfort when speaking to a person and not to a machine. The other advantage of the taped oral test was its ability to serve as an alternative to oral interview given situations when certified interviewers were not available (Lee, 2000). Other disadvantages included amount of time per interviewee and impracticality because of inconvenient geographical location (LaBlanc, 1997). Norris (1997) discussed the caveats of a Simulation German speaking test, such as: the examinees could communicate differently to exam-takers, characterizations of different language aspects for a tested level did not necessarily represent actual level ability of examinee. Variety of tasks and their specifications were also the focus of the studies. The scholars agreed that speaking tests benefited from having tasks on other language skills. A French Speaking test included reading along with oral response, scenario, and narration (based on the set of pictures) (LaBlanc, 1997).
In the review of recent research on speaking testing, Sandlund, Sundqvist, and Nyroos, (2016) examined 58 studies from 2004 till 2014 and identified main trends in this area. They pointed out an increase in other test formats of speaking exams along with OPI: paired and group tests. Unsurprisingly, a number of studies on comparison of formats appeared. The general consensus is that despite the obvious advantages of paired testing in terms of involving more interaction and better resembling natural conversation, the paired testing also presents a number of challenges in assessment of the performance and interaction between the examinees and raters (Brooks, 2009; Davis 2009; Ducasse & Brown, 2009; Birjandi, 2011). It should be noted that research on paired and group testing is still less common (Sandlund et al., 2106).

The array of facets of testing researchers are interested in is broad: examiner conduct, testing conditions, interactions, and task, to name a few. Simpson (2006) raised the question of differing expectations in speaking assessment. Deviation in expectations included variations in interlocutor behavior during the exam, divergent assumptions of language learners about the exam, and lack of agreement whether the exam should be conducted as an interview or a conversation. Planning task time is another feature of the test which relates to students' conduct, testing conditions, and task performance. Research shows evidence that planning time does not significantly affect the performance at the exam; though it should be considered as an option for the test takers (Wigglesworth & Elder, 2010).

While specific features of speaking exams get sufficient attention in the literature, relatively few studies (2004–2014) explored the sequential development of foreign language testing interaction (Sandlund et al., 2016). Furthermore, the literature on speaking tests and best practices in testing speaking in specific ESL programs is scarce. Practitioners would benefit from such action research in getting assistance and ideas on how to accommodate best practices in their particular education settings.

Problem statement: The existing standardized speaking tests target a general population of English learners without full consideration of the specific needs of certain groups of learners such as students in preparatory language schools. Those needs can be reflected in the gamut and number of tasks, topics, and scoring rubrics. The learners cannot be tested in all speech contexts. One critical issue is to evaluate the ability of the learner to successfully perform in future non-test contexts (Fulcher, 2003). The expectations and language should be connected to the content the students master during the academic year and measure the skills they obtain based on the material they studied. The ESL programs in preparatory schools at universities face a challenge to create an assessment tool which will allow them to efficiently assess large numbers of students who will enter a variety of departments upon completion and who have gone through an intensive year of language education.

RATIONALE FOR THE NEW FORMAT

Site

The School of Foreign Languages is in a southwestern Turkish university with English as a Medium of Instruction (EMI) at several faculties. The English Foreign Language Preparatory program at Muğla University has 47 instructors, including three instructors who focus on speaking skills, and an average of 450 students each year with 50 being foreign students. Although students in Turkey receive English education beginning in the second-grade of primary school, speaking and listening comprehension tend to present a major challenge for them in mastering English while in the preparatory school. The students are expected to be able to use appropriate language in general social and academic situations for informational and communication purposes. As a result, the overarching goal of our speaking classes is to promote English language competency for personal, social, educational and professional purposes, including applying English to real-life situations. Another goal is to help meet the needs of the faculties at the university and the Erasmus program. Those needs cover students' public speaking, in-class question production and critical thinking, project writing and presentation, small talk in English, absorbing and providing constructive criticism and feedback, proof-reading and summarizing. Thus, the expectation is that successful students leave the school with sharpened active learning and speaking skills, improved pronunciation, enhanced listening comprehension, and ease in conversation.

The previous format of the exam did not satisfy the expectations and the needs of all stakeholders and caused criticism from both students and faculty. The students were supposed to speak on two topics which they chose from two sets of monologue questions. Two assessors and one interlocutor had been used to administer exams. Later, this was changed to one interlocutor and one assessor. The assessors were rating the exam using an analytic rubric: each item (task completion, grammar, vocabulary, pronunciation, etc.) was evaluated separately. Considering our department's strained time schedule and no time for the detailed area-specific feedback, an analytic method appeared both inefficient and out of sync with how assessors grade students in practice. Itemized results were not shared with potential future departments, students or instructors. Due to time restraints, teachers also reported that rather than looking at each item, they were instead deciding on a score and then, sometimes arbitrarily, selecting the items needed to total that score – in essence forcing a holistic assessment onto an analytic scale. Also, in practice, when a sample of past graded exams were reviewed by a team of assessors, using both an analytic and holistic scale, similar results were achieved.

Bearing in mind all criticism and recognizing the need to make changes, the School administration made a decision to redesign the speaking exam practice. The goal was to develop a new format of the exam which would better meet the students' needs in the transition to their departments and the needs of instructors in evaluation of students' speaking skills. This exam would be more closely matched to the students' actual in-class speaking and learning experiences. The Speaking Unit (a subdivision of the EFL Test Office) comprised of foreign instructors was responsible for designing the test, which would: 1) allow for more objective evaluation of the students' ability to successfully perform not only in an academic environment but also in other speaking contexts; 2) align the tasks and structure with the content and materials the students were working with and studying during the academic year. Furthermore, the Speaking Unit was expected to provide necessary training for the interlocutors and assessors, and prepare the students for the exam in their speaking classes. The questions we asked included: Which mix and number of tasks do we choose? What is the topic complexity? What is the discourse type (descriptive vs. argumentative)? What is the setting? (De Jong et al., 2012; Fulcher, 2013).

Exam Structure

Our format for the current exam, based on instructor – student ratio, was to be one assessor and one interlocutor per student. Interlocutors would be allowed to give some feedback following the scoring from the assessor, also known as a rater, as necessary. Given this, how should we format the exam? "In terms of rating options, the best practice is to have multiple raters and multiple rating items. The next best practice is to have one overall evaluation item and multiple raters. In order of preference, the third choice would be to have one rater and multiple items. The least recommended solution would be to have one rater and one item" (Nakamura, 2004, p.51). For our situation, option 3 was chosen – one rater and multiple rating items. As it was also important to teach the students about the scale and how to measure their performance, it was designed to be adaptive to other speaking activities including projects and classroom activities in order to improve student familiarity.

The structure of the exam was thought to reflect the situations the students were going to encounter in real life: everyday conversations, narratives on the suggested topics, sharing opinions and inferences. As test designers we had to make decisions on the mix of tasks and the type of scoring rubric which would serve the purpose of the test (Fulcher, 2003, p. 86). We considered it important to embed a variety of tasks within the test to meet the future needs of students (Fulcher, 2003, p.

131): independent tasks where the examinees express an opinion on a given topic without any accompanying materials as well as integrated tasks with some prompts (e.g. pictures) (Iwashita et al., 2008).Thus, the following structure of the exam was agreed upon:

1. Warm-up
2. Conversation with the interlocutor
3. Exam card

The description of each stage of the exam follows:
-Warm up: This section is the introduction to the exam, including sign-in, checking identification, and an explanation of the exam structure. Greetings are also exchanged.
-Conversation: An everyday topic is suggested by the interlocutor (from the list of questions). Interlocutors inspire a conversation with follow-up questions based on student responses for a period of approximately 2 minutes. The goal is to make the conversation as natural as possible encouraging emotionally colored extended responses rather than plain answers to the questions.
Examples of the questions to start a conversation:

How was your weekend?
Did you watch the ______ game?
What's the best meal you've had recently?
How are your classes going?
Have you talked to your family recently? How are they?
Did you see the news about ________? What do you think about it?
Which city are you from? Tell me about it.
What's the best movie you've seen recently?
Any vacation plans? Which country/city would you like to visit?

-Exam Card: The tasks are aimed at evaluation of the test-taker's ability to talk for a minute and a half on a topic "displaying a range of functional and discourse skills such as description, argument, narration and speculation" (Brown, 2003, p.7).During this stage of the exam, interlocutors make very little interference. First, they make sure the examinee understands the questions or tasks. They can explain the unknown words if necessary in English. During the student response, the interlocutors are allowed to ask two prompt questions that encourage more speech without providing language assistance. We developed 49 cards in total (a card sample is presented in appendix A). Each card is comprised of 2 main components:

Monologue question

The topics are based on the units of the Speakout course book (Williams, 2015) "relevant to students' lives and global in nature" (p. 19). The topics for the monologue speech are presented in the form of situations or an invitation for discussion with prompts.

Are celebrities good role models for young people/adults? How can they use their influence in a good way?

It is now summer and school has ended. You and your friends are planning a vacation. Which country/city would you visit and why?
You win 2 tickets from a radio station, and can attend any music concert you want. Which singer will you choose to watch? Who will you take with you to the concert and why?
You have a test next month, and want to prepare yourself. What are the best techniques you will use to practice and improve your English speaking?
Your friend is a comedian, and is working on her next show. She asks you to tell her about something that made you laugh so that she can include it in her show. What would you tell her?
You have a new job and your boss wants to send you to another country. Would you go abroad and earn more money or would you stay and be close to your family and your friends. Why?

Picture description component
a. Photos cover a wide range of themes and issues: ecological problems, sport, leisure time, family life, food, emotions, social media, etc. The students are expected to describe what they see in the picture, draw inferences, and share their assumptions about what they see.
b. Response to a follow up question on a topic related to the picture.
Is it better for children to grow up in the countryside or in a big city? (a picture with an overcrowded metro station)
How can technology be beneficial/helpful for children? (a picture of a child playing with a tablet in a car)
How have people harmed or changed the environment? (a picture with a rescue worker saving a koala)

Rating rubric
After making the decision to redesign the exam, we had to decide how best to assess it. We administer, on average, 4 speaking exams per academic year. Time constraints in both giving students the opportunity to demonstrate their skills while also limiting burnout for assessors was a main concern. To properly do this, we needed to create a rubric that could help assessors efficiently determine the level of performance students had achieved in relation to what the speaking exam was attempting to measure.

According to Susan M. Brookhart (2013), “A rubric is a coherent set of criteria for students' work that includes descriptions of levels of performance quality on the criteria”(p.7). Appropriate well-thought out criteria and detailed descriptions of performance ensure the quality and effectiveness of the rubric. In this sense, we knew our criteria needed to be clear and specific and also that it needed to match the goals of our speaking program in preparing students for English based education in their future university departments.

A clear understanding of the purpose of our program and expected benchmarks for our students would allow us to clearly define the criteria of the rubric for each section of the speaking exam. However, the best way to assess and score these criteria was still a source of concern. Factors that affect the choice of a scoring approach include the required turnaround time for score reports, the qualifications of raters, and the number of qualified raters. The student population of the School of Foreign Languages at Muğla Sıtkı Koçman University continues to grow each year. Given the number of students and limited number of faculty, we needed a rubric that could provide high inter-rater reliability in a short period of time. In general, there are two main types of rubrics, holistic and analytic. Historically, our school has used analytic rubrics in all speaking and writing exams.

An analytic rubric is highly effective in helping teachers specifically identify areas of strength and weakness in a student's performance. After this, the feedback should be given to the student so they can understand it and make appropriate corrections. This type of rubric is also suitable when a student's performance needs to be shared with a program or department. Analytic rubrics are time consuming as they measure in detail different factors and criteria separately – for example, task completion, grammar, vocabulary use, pronunciation, etc. each with a separate score for each test item. However, they are very useful in terms of high-stakes measurement and corrective work. On the other hand, holistic style rubrics look at overall performance on a task or set of tasks. The criteria and targets for these tasks are still defined but unlike an analytic rubric, these criteria and targets are measured and observed as a whole rather than as individual assessment items. This creates a challenge for giving specific feedback and for weighting certain criteria over others but reduces time and allows for more consistency among assessors. In essence, an analytic rubric provides the fine details while a holistic rubric gives you an overall summarized view. A comparison of both scoring methods can be seen below.

Analytic scoring requires raters to determine whether specific characteristics or features are present or absent in a response.

It is useful if:

Differential weightings can be attached to different subscales, depending on the purpose of the assessment

Diagnostic information will be provided about test-takers' strengths and weaknesses

There are expectations of meaningful feedback

The drawbacks are as follows:

Time consuming to create and complete

Requires a strong and clear set of criterion and individual points of measurement for assessors to come to a similar score

Holistic scoring employs a scoring scale and training samples to guide raters in arriving at a single qualitative evaluation of the response as a whole.

The main features are as follows:

Speaking is not broken up into multiple subscales and weighting can only be done based on task.

One general speaking score is given based on a proficiency level that is outlined in detail.

There is an emphasis on what the learner is able to demonstrate, rather than what s/he cannot do.

Saves time by minimizing the number of decisions raters make and, with trained raters, it can be applied consistently and reliably

When student performance is at varying levels spanning the criteria, choosing the best descriptor can be hard.

With all things considered, for efficiency and accuracy, a holistic approach seemed to be the best. According to the Educational Testing Service (ETS), the following considerations should be at the core in developing scoring rubrics: the purpose of the assessment, the ability levels of the test-takers, and the demands of the task. The scoring rubric should be "aligned with the directions and task to ensure that raters are applying the appropriate scoring criteria and are not influenced by atypical response formats or the presence of extraneous information that could bias their scoring"(p.10). Care was taken to choose measureable criteria with clear directions for what the students should do and what the assessors should look for in three separate test items: Conversational English, Monologue, and Picture Description and Analysis. A variety of sample holistic rubrics were reviewed along with tested analytic rubrics from IELTS and TOEFL and a holistic rubric was developed with an analytic version for comparison.

In order to create detailed category performance descriptors, a model of criteria factors for scoring from Fairfax County Public Schools, one of the highest performing public school districts in the USA with a large foreign language program, was adapted. For the student assessment sheet, a 0-5 point scale for each test item was chosen to measure each performance descriptor and a half-point score range was given in each category to allow for nuance and flexibility. For example, a student can get a score of either 5 or 4.5 in the "Exceeds Expectations" category. For those instances where a student's performance might fall between two categories, a comparison is to be made between the two criteria and how many descriptors from each the students matches with the most. A maximum of 15 points can be scored on this exam, with a conversion scale for those instances that a 100% scale is needed. Additionally, separate rubrics for B1 and B2 level students were developed (Appendix B).

Instructor support

In a speaking examination, it is incumbent to standardize the procedure to a great degree for the interlocutor and the assessor to adhere to a framework in order to ensure reliability and fairness. It also aids in reducing diversity in administration of the exam, which can present different levels of challenges for exam takers (Brown, 2003, p.19). The test specifications we were looking at included settings, interactions involved (interlocutor, assessor, position, roles), communication events and activities (Fulcher, 2003, p. 123). Additional factors were considered in creating the guidebook and training for assessors. These issues included: how to deal with L1 usage to minimize bias with relation to foreign students, how to direct students and when to grade each item on the exam; how to deal with interruption of students' speaking. These things were done to reduce bias, increase reliability, and warrant consistency across testing environments.

For the purpose of our exam, we define the interlocutor as "a suitably qualified and trained person with whom a candidate interacts during a test in order to complete a speaking task. " (MANUAL on the English Language Proficiency Assessment: (ICAO language proficiency requirements). The principal function of the interlocutor is to ensure that the test takers have the opportunity to perform the various communicative tasks required by the test to the best of their ability and without undue stress.

The interlocutor is expected to:

Use standardised wording.
Elicit as much language as possible from the student.
Give instructions to the student.
Provide transition from one part of the exam to another.
Manage the interaction according to the instructions set out in the guideline.

Give the test takers adequate opportunities to speak.
Keep to the time limits.
Ensure that the interaction is recorded properly.
Know the format of the speaking examination very well.
Be familiar with all relevant instructions and requirements relating to the specific examination
Be familiar with the test formats and procedures

The major concern was adequacy of interlocutors' training (Brown, 2003). It was important to familiarize our teachers with the new format of the test and to describe the roles of the interlocutor. Making a guide for the interlocutors was an essential part for piloting the test. The purpose of this guide was to standardize the work of all the interlocutors. We collaborated on a teacher-friendly guide that could be effectively and easily used.

The new manual for the interlocutor was presented during a "Speaking Exam Workshop" as a training material for the future interlocutors. Sample questions and the new rubric were also reviewed and discussed during this workshop. All teachers were introduced to each part of the exam, standardised language, timing and expected behaviour as well as things that should or should not be done during the examination. They also had time to get familiar with the format and the manual itself.

All teachers received a copy with time to read and scan the material. During our first workshop teachers gave feedback and shared their concerns and observations about the manual. The comments were instrumental in improving the booklet.
The sections of the manual include:

Timing for each part of the exam
General description of the whole process and explanation of particular actions. In this part the interlocutor can see what to do, how and when to provide transition from one part of the exam to another.
Standardized language. Every part of the exam has its own appropriate language that unifies the spoken language of all interlocutors.
Goals of each section. We stated a set of goals we want students to reach in each section.
Every section of the examination is scrutinized in detail and clearly designed; so the interlocutor can easily follow it during the exam. In addition to the manual, every interlocutor has a very simple and clear to follow summary of the guide during the examination. It highlights the most important parts of the main guide, its extended version with many details.

OUTCOMES/CONCLUSIONS

The new format of the exam was piloted and recorded in the winter of 2016. It allowed for feedback from both students and instructors and provided data about what worked well and which aspects needed improvement. Following this exam, unofficial straw polls were taken from students as they exited. Upon review of a sample of videos to control for grading and admınistration of the exam, a follow-up training/discussion of the speaking exam was conducted. As a result of this pilot and the subsequent feedback, some exam questions and photos were edited or changed and instructions were amended. There was an overall positive response from assessors and interlocutors as well as from students on the flow and comfort of the exam. This was evidence of achievement of our goal of stress reduction for all participants.

Additionally, while scoring by initial assessors was generally the same or similar when compared to exams later reviewed via video by the test development team, there were some outliers. To increase consistency of results, an additional training was developed and reminders were sent to exam administrators prior to the exam.

The test specifications and materials are not "monolithic documents but working plans and definitions that should be frequently revisited" (Fulcher, 2003, p.136). As we continue to develop this new version of the speaking exam, efforts are regularly made to ensure materials remain current, new and veteran faculty alike are updated on training, and that the exam is applied as consistently as possible. To get more insight of strengths and weaknesses of our format we plan to use it with the students from our partnering university in Ukraine. It will allow us to identify the areas we should explore and improve. Planning time, assistance to students in preparation for the exam, and a possible shift to pair testing are on the research agenda of the Speaking Unit.

REFERENCES

American Council on the Teaching of Foreign Languages (1986) *ACTFL Proficiency Guidelines*, American Council on the Teaching of Foreign Languages, Hastings, Hudson, NY, USA.

Birjandi, P. (2011) *From face-to-face to paired oral proficiency interviews: the nut is yet to be cracked*, English Language Teaching, 4(2) 169–75.

Brookhart, S. M. (2013*)* How to Create and Use Rubrics for Formative Assessment and Grading, ASCD, Alexandria, VA, USA:

Brooks, L. (2009) *Interacting in pairs in a test of oral proficiency: co-constructing a better performance*, Language Testing, 26(3), 341–66.

Brown, A. (2003) *Interviewer variation and the co-construction of speaking proficiency*, Language Testing, 20(1), 1-25. doi: 10.1191/0265532203lt242oa

Council of Europe (2011) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Council of Europe.

De Jong, N.H., Steinel, M.P, Florijn, A., Schooner. and Hulstijn, J.H. (2012) *Facets of speaking proficiency*, Studies in Second Language Acquisition, 34 5-34.

Davis, L. (2009) *The influence of interlocutor proficiency in a paired oral assessment*. Language Testing, 26(3) 367–96.

Ducasse, Ana Maria and Brown, A. (2009) *Assessing paired orals: raters' orientation to interaction*. Language Testing, 26 (3) 423–43.

Fulcher, G. (2003) Testing Second Language Speaking, Pearson Education LTD, Great Britain

Iwashita, N., Brown, A., MCNamara, T., and O'Hagan, S. (2008) *Assessed Levels of Second Language Speaking Proficiency: How Distinct?* Applied Linguistics, 29(1) 24–49. doi:10.1093/applin/amm017

LeBlanc, L. B. (1997) *Testing French Teacher Certification Candidates for Speaking Ability: An Alternative to the OPI,* The French Review, 70(3) 383-394.

Lee, L. (2000) *Evaluating Intermediate Spanish Students' Speaking Skills through a Taped Test: A Pilot Study*, Hispania, 83(1) 127-138.

MANUAL on the English Language Proficiency Assessment: (ICAO language proficiency requirements) [online]. : 44 [cit. 2015-11-08]. DOI: MANUAL on the English Language

Mertler, C. A. (2001) *Designing scoring rubrics for your classroom,* Practical Assessment, Research & Evaluation, 7(25).

Nakamura, Y. (2004) A comparison of holistic and analytic scoring methods in the assessment of writing. The Interface Between Interlanguage, Pragmatics and Assessment, Proceedings of the 3rd Annual JALT Pan-SIG Conference.

Norris, J. M. (1997) *The German Speaking Test: Utility and Caveats*, Die Unterrichtspraxis / Teaching German, 30(2) 148-158.

Roca-Varela, María Luisa, and Ignacio, M. Palacios (2013) *How are spoken skills assessed in proficiency tests of general English as a foreign language? A preliminary survey.* International Journal of English Studies, 13(2) 53–68.

Sandlund, E., Sundqvist, P., and Nyroos, L. (2016) *Testing L2 talk: A review of empirical studies on second-language oral proficiency testing,* Language and Linguistics Compass, 10(1) 14–29. DOI: 10.1111/lnc3.12174

Simpson, J. (2006) *Differing expectations in the assessment of the speaking skills of ESOL learners,* Linguistics and Education, 17 40–55. DOI:10.1016/j.linged.2006.08.007

Stansfield, C. W., Kenyon, D. M., Doyle, F., Paiva, R., Ulsh, I., Cowles, M. A. (1990) *The Development and Validation of the Portuguese Speaking Test,* Hispania, 73(3) 641-651.

Trinity College London (2015) *Integrated Skills in English (ISE) Specifications -Speaking & Listening ISE Foundation to ISE III (First edition)*, Trinity College London, London.

Wigglesworth, G. and Elder, C. (2010) *An Investigation of the effectiveness and validity of planning time in speaking test tasks,* Language Assessment Quarterly, 7(1) 1-24. DOI: 10.1080/15434300903031779

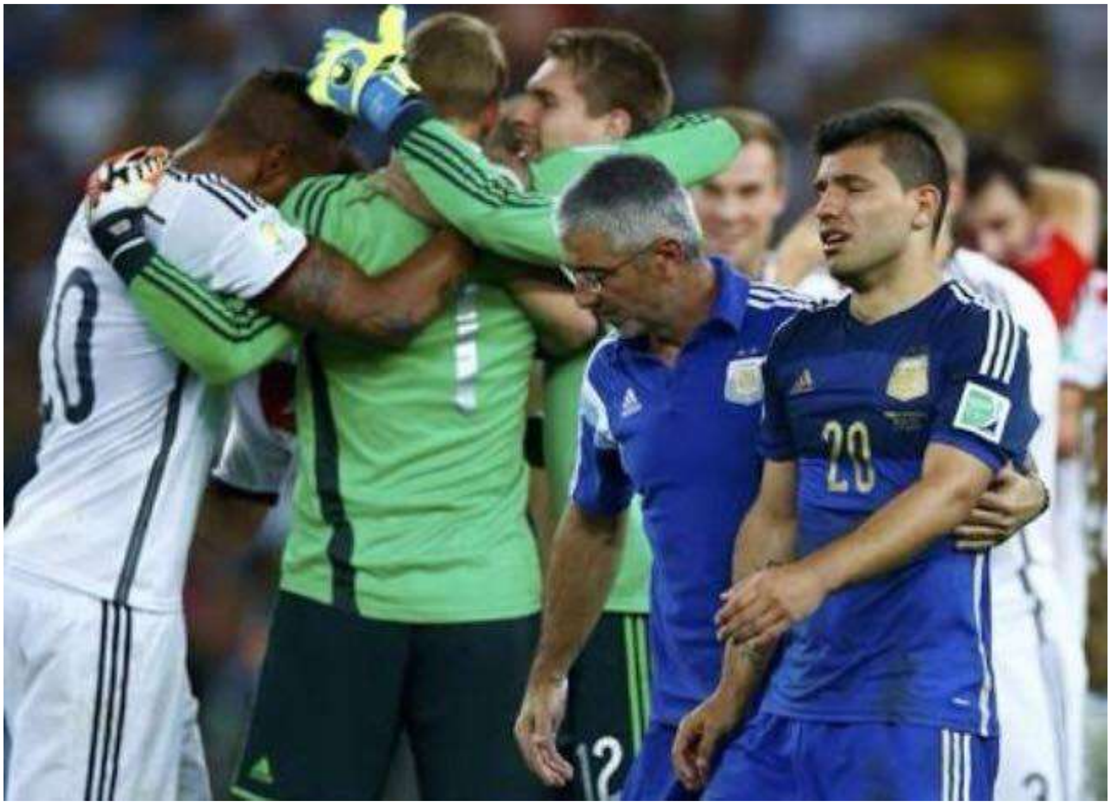
Williams, D. (2015) *Speakout. Intermediate teacher's book*. (2d Ed.). Pearson Education Limited.

Young, John W., Y. So and G.J. Ockey (2013) Educational Testing Service Guidelines for Best Test Development Practices to Ensure Validity and Fairness for International English Language Proficiency Assessments, 9-12.

Appendix A
Card 13
Monologue Question:

Think about your great, great grandchildren 100 years from now, will their world be better or worse than ours? Why, and in what ways would it have improved or worsened?

Picture Question:



How emotional do you get during sports matches/games?

Appendix B

Mugla Sıtkı Koçman University School of Foreign Languages
2015/2016 B2 Proficiency Exam: Speaking
Student Name: Exam Room:
Question Card: Session: Morning / Afternoon

	Exceeds Expectations	Meets Expectations	Almost Meets Expectations	Doesn't Meet Expectations	Basic/No Response
Task 1 Conversational English Responds to guided questions in a conversation about everyday events	5 – 4.5	4 – 3.5	3 – 2.5	2 – 1.5	1 - .5 - 0
Task 2 Monologue Provides an opinion/ experienced-based thorough response to a question that requires little to no verbal prompting from interlocutor	5 – 4.5	4 – 3.5	3 – 2.5	2 – 1.5	1 - .5 - 0
Task 3 Picture Description and Analysis Accurately details the characteristics and action of a picture using descriptive vocabulary Justifies interpretations Thoroughly answers the follow-up question with examples	5 – 4.5	4 – 3.5	3 – 2.5	2 – 1.5	1 - .5 - 0

Total Points _____

Guide for Assessors: If a student's overall performance matches the majority of the descriptors in a category, CIRCLE a choice in that category of the high or low end score. "Basic to No Response" can range from a one/two sentence response (1 point) to simple vocabulary words (.5 points) to a complete inability to respond (0). Add each section for the total score.

Exceeds Expectations Task Completion: Superior completion of the task; responses appropriate and with elaboration Comprehensibility: Responses readily comprehensible, requiring almost no interpretation on the part of the listener Fluency: Speech continuous with few pauses or stumbling Pronunciation: Enhances communication Vocabulary: Rich use of vocabulary Language Control: Control of advanced language structures with few or small errors	Meets Expectations Task Completion: Completion of the task; responses appropriate and adequately developed Comprehensibility: Responses comprehensible, requiring minimal interpretation on the part of the listener Fluency: Some hesitation but manages to continue and complete thoughts Pronunciation: Does not interfere with communication Vocabulary: Adequate and accurate use of vocabulary Language Control: Emerging control of advanced language structures, controlled use of basic language structures
Almost Meets Expectations Task Completion: Partial completion of the task; responses mostly appropriate yet undeveloped Comprehensibility: Responses mostly comprehensible, requiring interpretation on the part of the listener Fluency: Speech choppy and/or slow with frequent pauses; few or no incomplete thoughts Pronunciation: Occasionally interferes with communication Vocabulary: Somewhat inadequate and/or inaccurate use of vocabulary Language Control: Emerging control and use of basic language structures	Does Not Meet Expectations Task Completion: Minimal completion of the task and/or responses frequently inappropriate Comprehensibility: Responses barely comprehensible Fluency: Speech halting and uneven with communication Pronunciation: Frequently interferes with communication Vocabulary: Inadequate and/or inaccurate use of vocabulary Language Control: Inadequate and/or inaccurate use of basic language structures