REPUBLIC OF TURKEY
ÇAĞ UNIVERSITY
INSTITUTE OF SOCIAL SCIENCES
ENGLISH LANGUAGE TEACHING DEPARTMENT

AN ANALYSIS OF THE FINAL SPEAKING EXAM AT AN ENGLISH PREPARATORY SCHOOL IN TURKEY

THESIS BY
Özlem YILDIZ

SUPERVISOR
Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ

MASTER OF ARTS

MERSİN, March 2013

REPUBLIC OF TURKEY

ÇAĞ UNIVERSITY

DIRECTORSHIP OF THE INSTITUTE OF SOCIAL SCIENCE

We **certify** that this thesis under the title of **"AN ANALYSIS OF THE FINAL SPEAKING EXAM AT AN ENGLISH PREPARATORY SCHOOL IN TURKEY"** is satisfactory for the award of the degree of **Master of Arts** in the Department of **English Language Teaching.**

....

Supervisor – Head of Examining committee : Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ

....

Member of examining committee : Assist. Prof. Dr. Hülya YUMRU

Kim Raymond Humiston

....

Member of examining committee : Assist. Prof. Dr. Kim Reymond HUMISTON

I certify that the signatures belong to the above-named academicians.

T.C. ★ ÇAĞ ÜNİVERSİTESİ SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜ 1923

01/03/2013

Assoc. Prof. Dr. Haluk KORKMAZYÜREK

Director of the Institute of Social Sciences

Note: The uncited usage of the reports, charts, figures, and photographs in this dissertation, whether original or quoted for mother sources, is subjected to the Law of Work of Art and Thought No: 5846

ÖZET

BİR TÜRK ÜNİVERSİTESİ HAZIRLIK OKULU'NUN KONUŞMA BECERİSİ FİNAL SINAVI DEĞERLENDİRMESİ

Özlem YILDIZ
Yüksek Lisans Tezi, İngiliz Dili Eğitimi Anabilim Dalı
Tez Danışmanı: Doç. Dr. Şehnaz ŞAHİNKARAKAŞ
Mart 2013, 88 sayfa

Bir dil becerisi olarak konuşma, her zaman yabancı dil eğitimcileri ve öğrenenleri için önemli bir yere sahiptir. Bunun sebebi belki de bir yabancı dili konuşmadaki başarının o yabancı dildeki genel başarı için bir gösterge olarak kabul edildiği gerçeğidir. Konuşma becerisi en önemli dil becerilerinden biri olarak kabul edilir, ancak değerlendirilmesi de en zor olan beceridir. Bu nedenle en uygun konuşma becerisi değerlendirilme yolları birçok çalışmaya konu olmuştur.

Bu çalışmanın ana amacı Zirve Üniveristesi Hazırlık Okulu'nda uygulanan konuşma becerisi final sınavının problemli noktalarının olup olmadığının tespit edilmesidir. Bu amaca ulaşmak için sınavın içerik geçerliliği ve sınavı değerlendirenler arasındaki tutarlılık incelenmiştir. Bunlara ek olarak, öğrenci ve öğretim elemanları ile yapılan görüşmeler sonucunda sınav hakkındaki görüşler ve öneriler elde edilmiştir.

Çalışmaya Zirve Üniversitesi Hazırlık Okulu'dan 15 öğrenci ve 10 öğretim elemanı katılmıştır. Katılımcılarla yapılan görüşmeler sonucu elde edilen veriler sınavın zayıf noktalarının tespit edilmesi ve bu noktaların iyileştirilmesi için öneri getirmek üzere kullanılmıştır.

Çalışma sonucunda konuşma becerisi final sınavının içerik geçerliliği vasfına sahip olmadığı saptanırken, değerlendiriciler arasındaki tutarlılığın ise yüksek olduğu belirlenmiştir. Ayrıca, öğretim elemanları ve öğrencilerle yapılan görüşmeler sonucu elde edilen veriler kullanılarak Zirve Üniversitesi Hazırlık Okulu konuşma becerisi final sınavının iyileştirilmesi için önerilerde bulunulmuştur.

Anahtar Kelimeler: Değerlendirme, Konuşma Becerisi Değerlendirme, Geçerlilik, Güvenilirlik

ABSTRACT

AN ANALYSIS OF THE FINAL SPEAKING EXAM AT AN ENGLISH PREPARATORY SCHOOL IN TURKEY

Özlem YILDIZ

Master of Arts, English Language Teaching

Supervisor: Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ

March 2013, 88 pages

Speaking as a language skill has always been important both for foreign language educators and learners. The reason for this may be the fact that the success in speaking a foreign language is considered as an indicator for success in that language. However, although speaking ability is one of the most important skills, it is also the most difficult skill to assess. Therefore, appropriate ways for assessing speaking skill has been one of the most common research issues in this field.

The main purpose of this study is to investigate whether or not the speaking component of final exam at Zirve University Preparatory School has any problematic areas. In order to reach this aim content validity and inter-rater reliability of the exam are examined. Moreover, the ideas and suggestions about the exam were gained by interviews held with the students and the instructors.

The participants of the study are 15 students and 10 teachers of Zirve University Preparatory School. Interviews conducted with the participants are used to gather data for identifying the weak points of the exam and suggesting appropriate and better ways to do the speaking exam. It was found out that although the exam was discovered not to have content validity, the inter-rater reliability of the exam was satisfactory. Furthermore, using the participants' opinions and recommendations gathered from the interviews, ways for a better speaking assessment procedure for Zirve University Preparatory School was suggested.

Keywords: Assessment, Speaking Assessment, Validity, Reliability

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my thesis advisor, Şehnaz Şahinkarakaş, for her invaluable guidance, patience, encouraging and understanding attitude. I've learnt a lot from her and always felt lucky to have the privilege to study with her.

I owe my special thanks to my family, especially my mother, Meliha Yıldız, who always supported and motivated me throughout my study and my life. My sister Özgen Yıldız and brothers Özkan Yıldız and Hakan Yıldız for their energy, enthusiasm and trust. Their existence is a reason good enough to struggle and continue. Without their trust and belief I will not be able to feel the strength to proceed this study.

And, finally, I would also like to express my profound love and greatest thanks to my beloved friend Tuba Arman, for being an indispensable part of my life, and the half of my heart and soul. She has always been with me during my hardest times with her constant support, love and encouragement. Without her valuable contributions and encouragement, this study wouldn't have been possible.

01 MART 2013
Özlem YILDIZ

ABBREVIATIONS

ELT	:	English Language Teaching
CEFR	:	Common European Framework
CLT	:	Communicative Language Teaching
L1	:	First Language
L2	:	Second Language
SPSS	:	Statistical Package for Social Scientists

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 1

1. INTRODUCTION

This chapter contains six sections. The first section is the background of the study. Then statement of the problem, purpose of the study and research questions are given. Lastly, significance of the study and operational definitions are pointed out.

1.1. Background of the Study

Speaking is an integral part of people's daily lives. Speaking as a skill has a fundamental place among other skills as most of the communication depends on speaking skill. For many people, learning a second language means learning to speak in that language and success is measured by how well one can carry out a conversation in that language (Nunan, 2002). For this reason, speaking skill is of vital importance in terms of language learning and teaching.

Although being able to speak a language is regarded as mastering that language, speaking as a language skill has been the most neglected one. We can witness this kind of laxness in language teaching methods throughout the history as well. Learning and teaching grammar of a language used to be considered as the most important aspect, and the focus was on structure, language patterns and memorization. In 1960s, with the rise of communicative approach, the role of speaking ability has become more important in language teaching, and language started to be considered as a tool for communication.

Speaking is also an important part of the curriculum in language teaching which makes it an important object for assessment as well. As a result of Communicative Language Teaching (CLT), performance testing, especially testing the speaking ability has become one of the important issues in language testing (Sak, 2008). Because of the nature of speaking ability, assessing speaking is challenging and there are many limitations in this area. According to Luoma (2004), there may be some inconsistencies in the evaluation process as speaking requires a candidate to use language in some way due to its interactive nature.

There are many practical restrictions of this area like administrative costs, difficulties of testing a large number of students either individually or in small groups, training the examiners and the total amount of time and the number of examiners needed for administering the tests (Sak, 2008). In addition to all these, the most important aspect of testing speaking may be that scoring requires human raters and as there are many factors that can influence our impression of how well someone can speak a language, the results can inevitably be subjective. Brown (2005) highlights the problem as follows "the subjective nature of the scoring procedures can lead to evaluator inconsistencies or shifts having an affect on students' scores and affect the scorer reliability adversely" (p.191).

Although there are such kind of restrictions, there are different ways that institutions use to assess oral performance of the students somehow. This being the case, arriving at the most appropriate, valid and reliable decisions about the test construction, procedure and scoring of a speaking exam process is probably the most challenging task that an institution has to carry out.

1.2. Statement of the Problem

Students generally tend to experience difficulty in developing oral fluency. This is mostly because of the fact that speaking as a language skill does not usually gain due importance in Foreign Language Education in Turkey. In addition to this, one can not ignore the fact that assessing speaking is as difficult as teaching it. As speaking assessment requires many factors to be taken into consideration, it is the most troubled part of assessment process at Zirve University Prep School as well.

Since preparing and administering valid and reliable speaking exams is a considerably difficult and complex work, and the rest is much easier; there still has not been a standardized procedure for speaking assessment at Zirve University Prep School. Although different ways of speaking assessment have been tried so far by the institution, no certain procedure has been decided on yet. Therefore, the weaknesses and strengths of the speaking component of the final exam are needed to be identified and certain special steps are needed to be taken to ensure a much more reliable and valid exam procedure at Zirve University Prep School.

1.3. Purpose of the Study

This study aims to identify the problematic aspects of the speaking assessment procedure and the difficulties that the instructors and students face at Zirve University Prep School. The students' and the instructors' ideas and suggestions are also tried to be discovered in order to use to solve these problems. Another purpose of the study is to investigate the reliability and validity of the speaking component of the final exam, and in the light of these findings suggesting a speaking assessment procedure which would be appropriate for Zirve University Prep School is the eventual goal of this study.

1.4. Research Questions

Considering the aims stated above, this study intends to answer the following research questions regarding speaking assessment procedure at Zirve University Prep School.

R.Q. 1: How valid is the speaking component of the final exam at Zirve University Prep School?

R.Q. 2: How reliable is the speaking component of the final exam at Zirve University Prep School?

R.Q. 3: What do the instructors think about the speaking assessment process?

R.Q. 4: What do the students think about the speaking assessment process?

R.Q. 5: What would be the most appropriate speaking assessment procedure at Zirve University?

1.5. Significance of the Study

Speaking is the most difficult language skill to assess reliably. Many reasons can be stated for this problem. As we have to make instant judgements about a person's speaking ability during a face-to-face interaction and many different factors such as "many features of speech (pronunciation, fluency, accuracy and so on), language level, gender, status and personal characteristics of the interlocutor, questions asked, tasks presented and the opportunities that are provided" are all playing roles in this judgement process, assessing speaking is not impossible, but difficult (Luoma, 2004, p. 21).

When we look from this perspective, this study is significant for some reasons. First of all, speaking assessment has an important role in the evaluation process at Zirve University Prep School. Since the students are required to take speaking exams 8 times in an education year, these exams play an important role in the decisions of students' overall performances, which say they either pass or fail the program. This reveals the need for evaluating these exams in terms of validity and reliability.

Secondly, the findings of this study will be useful for the instructors, administration and test constructers at Zirve University Prep School and will reveal the problematic aspects of speaking exam procedure and the reasons for lamenesses during this process as well. These will indicate for the institution what to change and what to keep about the exam process. What else makes this study important is that, the aim of the study is not only to identify the problems, but also to find out the students' and the instructors' ideas about the speaking exam process and show the situation from their point of view.

Moreover, since the findings will shed light on how well these exams are assessing the oral performances of the students, whether there is a need for making changes about the exam or following a different process will be determined. If there is such a necessity, with the help of these results, a better and more valid speaking exam procedure will be recommended for the institution to use.

Lastly, this research study and the findings of the study will be beneficial for the research area and the ones who want to study in this field. It will also be valuable for other institutions with similar problems, test constructors and administrators in terms of revealing better and more correct ways of assessing speaking.

1.6. Operational Definitions

Assessment: Assessment is the systematic collection, review and use of information about educational programs to improve student learning. Assessment focuses on what students know, what they are able to do, and what values they have when they graduate. Assessment is concerned with the collective impact of a program on student learning (Hughes, 2003).

Speaking Assessment: Speaking is the most difficult language skill to assess reliably. A person's speaking ability is usually judged during a face-to-face interaction, in real time, between an interlocutor and a candidate. The assessor has to make instant judgments about a wide range of aspects of what is being said (Bachman, 2004).

Validity: It is considered to be the degree to which a test measures what it is supposed to measure (Hager, et al.,1994).

Reliability: It is the consistency of evaluation of results.

Interlocutor: It refers to a separate person who interacts with the candidate while the examiner assesses the performance (Alderson, et al., 2005).

Grader: This word, which is also named like assessor, examiner or rater, indicates the person who is responsible for judging a candidate's performance in a test (Alderson, et al., 2005).

CHAPTER 2

2. REVIEW OF LITERATURE

In this chapter speaking as a skill and speaking types, problems of testing speaking are explained. In the second part of the chapter the qualities of the tests like validity and reliability are reviewed. Content validity, face validity, criterion-related validity and validity in scoring are discussed in detail. Inter-rater reliability and intra-rater reliability are identified later on. Additionally, speaking task types and speaking test scales are presented. Rater-interlocutor training is the last part.

2.1. Speaking as a Language Skill

Communication competence can be defined as “the ability to use language to communicate within a specific situation” (Rubin, 1982, p.19). Thus, our main aim in foreign language teaching is to have our students be able to use the target language. In the first place it may seem to be adequate to teach grammar and vocabulary, but in order to ‘use’ a language these will not be enough. Bygate (2001, cited in Güney, 2010) makes a comparison between driving a car and using a language, by saying that both requires knowledge at first, practice is needed for being proficient, though. Taylor (2003) states that in order to name a speaker ‘proficient’ of speaking a language he needs to posses the following competences:

a) a wide repertoire of lexis and grammar to enable flexible, appropriate, precise construction of utterances in ‘real time’ (the knowledge factor);

b) a set of established procedures for pronunciation and lexico-grammar, and a set of established ‘chunks‘ of language, all of which will enable fluent performance with ‘on-line’ planning reduced to acceptable amounts and timing (the processing factor) (p. 2).

Speaking is regarded as the most difficult skill to be taught and assessed as it depends on the production of the language. Brown (2001) lists the reasons of this difficulty as: reduced forms of the language like contractions, elisions, reduced vowels, and so on; colloquial language like idioms or some certain phrases and features of language like stress, rhythm, intonation. According to Taylor (2003), “spoken language production tends to be based in social interaction, to be purposeful and goal-oriented within a context; and while it is capable of being routine and predictable, it also has the capacity for relative creativity and unpredictability” (p. 2). Speaking requires the language user to be simultaneous in many different abilities of the language, each of which develops at different rates and qualities. That is to say, a proficient speaker is accepted to be able to make use of pronunciation, grammar, vocabulary, fluency and comprehension at the same time in order to enable communication. Since it is quite difficult for a language learner to become able to achieve all these at the same time, speaking is commonly regarded as the most difficult language skill.

2.2. Types of Speaking

According to Brown (2001), it is possible to talk about five different types of speaking performance:

Imitative speech occupies a very limited part of speaking performance. It includes mostly drilling and it is not for a meaningful interaction, but for focusing on some particular element of language form.

Intensive speech "is designed to practice some phonological or grammatical aspect of language." Directed response tasks, reading aloud, sentence completion can be given as some examples for this kind of assessment types (Önal, 2010, p.10).

Responsive speech includes assessment tasks which are like simple requests and comments, standard greetings and small talk.

Interactive speech which is an extended form of responsive speech, is carried out for the purpose of conveying or exchanging specific information.

Extensive speech, which is also named as monologues, can be planned or spontaneous. Advanced level learners are expected to give monologues like oral reports, presentations, story-telling or summaries.

2.3. Problems of Testing Speaking

It is commonly accepted that speaking is the most difficult skill to test. There are many reasons that can be stated for this, but the most important thing is that it is a difficult construct to define. According to Kitao&Kitao (1996) "it involves a combination of skills that may have no correlation with each other, and which do not let themselves well to objective testing. There are not yet good answers to questions about the criteria for testing these skills and the weighting of these factors " (p. 2).

There can be many kinds of categorization of speech; it can be broken down into pronunciation, intonation, accuracy, fluency; or it can be categorized in terms of strategies; or it can be considered as a way of interaction and categorized using the methods of pragmatics or discourse analysis (Fulcher, 2003). As these standards are all important in daily usage as a whole it is not completely correct to separate out them which makes the test construction problematic. This means that "the accurate speaker may communicate slowly, whereas the fluent speaker may sacrifice accuracy for the sake of rapid communication" (Skehan, 1998, p. 24).

Another difficulty is separating the listening skill from speaking skill. Communication involves both listening and speaking at the same time, and inevitably speaking depends on comprehending spoken input which shows that there is an interchange between speaking and listening. This appears as another difficulty in testing speaking as you can not be completely sure that you are testing purely speaking or speaking and listening together (Kitao&Kitao, 1996).

Also success in speaking depends not only on the speaker but also the listener as well, since the degree to which the listener is familiar with the speaker's accent and background affect the communication situation. Heaton (1990) states his idea about this situation like:

> ..success in communication often depends as much on the listener as on the speaker: a particular listener may have a better ability to decode the foreign speaker's message or may share a common nexus of ideas with him or her, thereby making communication simpler. Two native speakers will not always, therefore, experience the same degree of difficulty in understanding the foreign speaker (p.88).

Furthermore, assessing speaking ability is difficult in terms of its being evaluated by human raters. Being highly subjective is one of its characteristics and according to Ur (1996), the most significant problem of testing speaking is reliability since there can be various judgements in assessing one's performance. For this reason, some inconsistencies about raters and scores are indispensable if some required procedures about the speaking exam like evaluating rater reliability, training raters and designing correct rating scales have not been done before the exam (Sak, 2008).

In addition to all these, there are also many problems about administering speaking exams especially when it is necessary to test large number of students. The number of students indicates how much time is needed as well, and this too much time sometimes becomes too impossible for some institutions to handle with a few teachers. The necessary number of examiners and raters, resources and preparation necessary for training these people, the amount of time needed, equipments and administrative costs and many other kinds of practical factors makes it nearly impossible to test speaking (Cohen, 1980; Weir, 1990).

2.4. Qualities of Language Tests

2.4.1. Validity

Validity is one of the most important features of good tests and has been a remarkable aspect of test construction. As Luoma (2004, p.184) defines “validity refers to the meaningfulness of the scores.” This is the core of the issue of test development as the scores do not mean what they are believed to mean, if a test is not valid for the purpose for which it is designed (Alderson, et al., 2005). Henning (1987, cited in Alderson, et al., 2005) gives the definition for validity as “the appropriateness of a given test or any kind of its component parts as a measure of what it is purported to measure” (p. 79).

2.4.1.1. Content Validity

Content validity is about how sufficient and proper a test is for measuring what it is supposed to measure. As Hughes (2010, p. 41) defines “ a test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned.” There should be a parallelism between the aim of the test and the test itself. The test would be accepted to have content validity only if it included a proper sample of the relevant structures (Hughes, 2010). The required judgements for content validity should be made by ‘experts‘ in a systematic way. According to Alderson et al. (2005), a common way to analyze the content of a test is to compare it with a statement of what the content ought to be.

For a test to be named as ‘accurate‘ it should have content validity, thus content validation is required to be carried out by the test developers before a test is being used.

2.4.1.2. Face Validity

According to Ingram (1977, cited in Alderson et al., 2005) face validity refers to the test’s “surface credibility or public acceptability”. It is completely related to the appearance of the test and commonly regarded as unscientific since it includes judgements of people who are not ‘expert’ like students, colleagues or administrators. As Hughes (2010) states, if such kind of a test is used, the test takers' reaction to it may mean that they do not perform on it in a way that truly reflects their ability.

2.4.1.3. Criterion-related Validity

The concept of criterion related validity involves ”demonstrating validity by showing that the scores on the test being validated correlate highly with some other, well-respected measure of the same construct" (Brown, 2005, p. 233). According to Hughes (2010), it refers to "the degree to which results on the test agree with those provided by some independent and highly dependable assessment of candidate's ability. This independent assessment is thus the criterion measure against which the test is validated" (p. 27). This so called independent assessment can be a similar or parallel version of the test or candidates' self assessment of their ability.

2.4.1.4. Validity in Scoring

In order to define a test as valid, validation of the test items will not be enough. Hughes (2010) makes the issue clear by saying that "it is no use having excellent items if they are scored invalidly" (p. 32-33). While scoring the responses in a reading test, considering grammatical problems as mistakes prove that the scoring is not valid. Therefore, instructions that the test developers will give gain importance in terms of scoring validity.

2.4.2 Reliability

Reliability is another important issue to be taken into consideration about testing. It must be accepted that we will not be able to get exactly the same results from a measurement when it is done twice, as it is contradictory to human nature to behave in the same way in a situation which is exactly the same as the previous. As this is the case, it is significant to remember that "reliability is not measured; it is estimated" (Sak, 2008, p.21). Thus, we will not expect the scores will exactly be the same, but similar. "The more similar the scores would have been, the more reliable the test is said to be" (Hughes, 2010, p. 36).

We quantify the reliability of a test in the form of a reliability coefficient, which allows us to compare the reliability of different tests (Nakamura, 1996). The ideal reliability coefficient is 1, which indicates the same results for two different testing situations of the same test.

Another issue to keep in mind about reliability is that a test cannot be valid unless it is reliable, because if a test does not measure consistently we can not expect that test to be measuring precisely (Sak, 2008).

2.4.2.1. Inter-rater Reliability

When the scoring requires no judgement the test becomes more objective, but we can not expect the testing situations which requires judgement like speaking performance of a test taker in an interview to be as objective as a multiple choice test. According to Sak's definition (2008), ”this type of reliability is estimated by examining the scores of two raters and calculating the correlation coefficient between the two sets of scores” (p. 23). Mead (1980) also mentions that “there are many forms of reliability; however, the most important form of reliability for a speaking measure is inter-rater reliability” (p. 2).

2.4.2.2. Intra-rater Reliability

Intra-rater reliability is estimated by gathering two sets of scores given by the same rater in different times for the same group of test takers. Weir (2005) explains this as: “each marker needs to be consistent within himself, i.e., given a particular quality of performance, he needs to award the same mark whenever this quality appears“ (p. 34).

2.5. Speaking Tasks

Task design is one of the most important elements in testing speaking as we somehow lead the test takers' talk by the tasks we give to them. The style of language we choose to use differs according to our purpose in a communication; similarly, the content and the format of the talk to be assessed will be guided by these tasks. The methods and tasks used for assessing oral communication skills should be chosen in accordance with the purpose of the assessment. A method that is appropriate only for giving feedback to students, is not appropriate for evaluating students at the end of a course.

Tasks can generally be defined as the activities that are used by a learner for a communicative purpose. Bachman and Palmer (1996) defined the term like: “Speaking tasks can be seen as activities that involve speakers in using language for the purpose of achieving a particular goal or objective in a particular speaking situation” (p. 89).

Speaking exams traditionally consists of meetings between a test taker and an examiner which is called one-to-one testing, but such kind of exam settings can be various according to the tasks we are using. A speaking exam can consist of pair or group tasks as well as individual tasks. The thing to be considered while designing tasks is that they must fit the aims of the assessment and represent the abilities we wish to measure. From this perspective, it is better to use various methods to collect information about students' speaking ability, instead of repeating the same tasks again and again. As Luoma (2004) states: "if someone is good at describing, it does not automatically mean that he or she is also good at comparing things, telling stories or justifying an opinion" (p. 32). This shows that a comprehensive speaking exam is better to include various series of short tasks.

2.5.1. Speaking Task Types

There are many kinds of speaking tasks to be defined and many kinds of definitions for these tasks as well. Burgess and Head (2005) categorized speaking tasks into four:

- Interview Tasks
- Presentation Tasks
- Negotiation Tasks
- Discussion Tasks

On the other hand, Luoma (2004) discusses the task types under two main headings, open-ended tasks and structured tasks. Open-ended tasks which generally require a stretch of talk, either with turns between speakers or a single long speaking turn, guide the speaking process but allow the test takers for different ways of achieving the task requirements (Luoma, 2004). Description, narrative, instruction, comparison, explanation, justification, prediction and decision tasks can be examples of discourse types of open-ended speaking tasks. Structured speaking tasks, conversely, "specify quite precisely what the examinees should say" (Luoma, 2004, p. 139). Such kind of tasks like reading aloud, sentence repetition, mimicry, factual short-answer questions, reacting to phrases contain limited production.

Descriptive tasks are one of the many common task types in speaking tests. According to Luoma (2004), they can be used in one-to-one interviews and with pairs and indicates how well the examinees can comprehensibly describe something they know. While developing such kind of tasks deciding the limits for parallel tasks is important. In case of using pictures, as the pictures we are using control the content, size and length of the speaking, developers need to be careful.

Narrative tasks show the examinees' ability to recount a sequence of events which are usually based on picture sequences. Asking examinees to tell something happened to them can be another alternative. Choosing the appropriate pictures is important for this type as well, since "they should generate enough talk and provide opportunities for the examinees to show what they know." (Luoma, 2004, p. 144).

The aim for *instruction* tasks is getting the message across; that's why it tends to be like short exchanges between the speaker and the listener. Instruction giving tasks suit live testing situations well (Luoma, 2004).

Comparing and contrasting tasks are generally considered more demanding than the others, as they require analysis and the discussion of similarities and differences with complex grammatical structures. Such kind of tasks can be based on not only pictures, but also concepts like comparing/contrasting urban life with rural life (Luoma, 2004).

According to Luoma (2004), *explaining and predicting* tasks, such as explaining the contents of a graph or explaining a process, are fairly common tasks. "To do well on the task the speakers need to set the scene and identify parts of the information or stages in the process that they are explaining and present them in a coherent order" (Luoma, 2004, p. 149).

Decision tasks require making a decision by discussing the issue that the decision concerns from a number of perspectives. Test developers should be careful about choosing the issues. As the speakers are asked to express and justify their opinions while stating arguments for and against, the issues to be discussed should not be clear-cut (Luoma, 2004).

2.6. Speaking Test Scales

While evaluating a speaking performance we come through scores that show us how well the examinees can speak the language being tested by using some criteria which are indicated mostly by numbers. According to Luoma (2004), "one way to elicit the construct of speaking ability for a certain context is through a scoring rubric which informs test users what a test aims to measure" (p. 54). Rating scales are the instruments that express the understanding of how good performances differ from weak ones and can be defined as the series of statements that describe what each score means from lowest to highest.

In order to ensure validity and reliability of a speaking performance test, we need to be careful about the scoring that is based on criteria specific to that particular testing context. McNamara (1996) states " a scoring rubric can affect the speaking assessment, as there may be an interaction effect between the rating criteria and the examinees' performance" (p. 36). Rating scales describe the kinds of speaking skills that the tasks require, so there is a direct connection between scales and the tasks used. Therefore, careful examination of how rating scales have a relation with speaking performance should be considered in order to decide the fairness of the speaking assessment (Kim, 2006).

2.6.1. Holistic Scales

According to Hughes (2010), this type of scales, which are sometimes referred to as impressionistic scoring, involve the assignments of a single score to a performance on the basis of an overall impression of it. Since the raters are expected to give only one score, these kinds of scales are practical for decision making. In addition, the greatest advantage of a holistic scale is its simplicity and speed (O'Sullivan, 2008). Its flexibility in allowing many different combinations of strengths and weaknesses within a level is named another advantage by Luoma (2004).

On the other hand, O'Sullivan (2008) also states some disadvantages of such kind of scales as follows:

> ... the disadvantage of this scale include the danger of 'trial by first impression' meaning that since the examiner is asked to give one score only he or she may (and often does) simply rely on their first impression (or previous knowledge) of the candidate. So the score awarded may not actually reflect the observed performance (p. 21).

2.6.2. Analytic Scales

This kind of scales aim to capture the examinees' performance on various aspects of communication. Luoma (2004) defines them as: "Analytic scales contain a number of criteria, each of which has descriptors at the different levels of the scale. The scales forms a grid, and the examinees usually get a profile of scores, one for each of the criteria" (p.68). Since they require a separate score for each of a number of aspects of the performance, there exist more details about the test takers' speaking ability.

Hughes (2010) lists many advantages of such kind of scales. First, they reveal the problems of sub skills in individuals as they contain more details. Secondly, these scales put the raters in a position that they have to consider many aspects of performance which they might otherwise ignore. Thirdly, as the raters have to give a number of scores, the test generally tend to be more reliable.

In contrast with these factors, the time that it takes to use these scales stands as a disadvantage. Additionally, since there are many details to be concentrated on, diverting attention from the overall effect of the performance seems inevitable; which can be stated as another disadvantage. “As the whole is often greater than the sum of its parts, a composite score may be very reliable but not valid.” (Hughes, 2010, p. 62).

2.7. Rater-Interlocutor Training

Speaking as a skill cannot be thought apart from listening. As such kind of tests somehow include listening as well, and as the test-takers are not the only people who produce talk during a speaking test, the interlocutor talk gains importance. Without a standardization, there is no guarantee that all interlocutors will ask the same questions in the same manner, which can make an important difference in students’ performances, and so the scores as well (Caban, 2003). Taylor (2003), also highlighted the problems of variation in interlocutor talk and its effects on the amount of the speaking opportunities that a candidate has, the candidate’s performance and the scoring for that performance, and suggests the use of a ‘standardized script’ or ‘interlocutor frame’ in speaking tests.

Salaberry (2000), states that the difference in the assessment criteria can show inconsonant results and says “while one rater may focus on pronunciation accuracy, another may find vocabulary to be the most salient feature” (p. 6). If the marking systems are not clear and there is a variety of understanding, interpretation and application of these marking systems, the scoring will inevitably be invalid. In order to eliminate such kind of inconsistencies rater and interlocutor training with a detailed set of assessment criteria can do considerably much. According to Alderson, et al. (2005, p. 105), “If the marking of a test is not valid and reliable then all of the other work undertaken earlier to construct a ‘quality’ instrument will have been a waste of time.” As a result, the training of exam staff is a crucial component of any testing program.

In addition to the inexactness of the assessment criteria, raters can also be biased towards specific candidates, topics, tasks, L1, academic backgrounds, socioeconomic status, race, gender and so on. Rater bias is one of the most serious problems with the evaluation process of a speaking performance. Sitinggs et al. (1985, cited in Carlson & Howell, 1995) states that “rater bias is a function of social experiences and attitudes of the rater, clarity and precision of the scoring criteria and standards and the extend to which the scorer has internalized those standards” (p. 136).

In order to alleviate the problems and negative effects of rater bias, Carlson and Howell (1995) emphasize the importance of rater training and say that many studies indicate training of raters improves the evaluation process. As the rater rather than the instrument is the most significant source of error, rater and interlocutor training seems to have crucial importance.

CHAPTER 3

3. METHODOLOGY

In this chapter, the design of the study is explained firstly. Then, the participants of the study, instruments which are used in the study and data collection procedures are explained. Lastly, some information about data analysis is provided.

3.1. The Design of the Study

This study was designed to find out whether the speaking component of the final exam at Zirve University Prep School has any problematic areas or not by investigating the reliability and validity of the exam, and the thoughts of the instructors and students who took part in the exam. In order to reach this aim, different kinds of instruments were used for the study. Thus, this study is both a qualitative and quantitative research study.

In the quantitative part of the study, in order to find out the inter-rater reliability of the final speaking exam, speaking scores of Level B students are used. This data was used to examine the Research Question 2 (*How reliable is the speaking component of the final exam at Zirve University Prep School?*). In the qualitative part of the study, interviews were held both with instructors and with students in order to find answers for the Research Questions 3 (*What do the instructors think about speaking assessment process?*), 4 (*What do the students think about speaking assessment process?*) and 5 (*What are the suggestions of instructors and students for a better speaking assessment procedure?*). Moreover, in order to examine Research Question 1 (*How valid is the speaking component of the final exam at Zirve University Prep School?*), a content validity interview was also conducted with the Listening and Speaking Coordinator of the English Language program.

3.2. Participants

3.2.1. Students

The students are used to gather data about the final speaking exam. The students participating this research study were chosen among Level B preparatory class students. This group involves 15 students who registered the preparatory program in 2012-2013 academic year. Six of them are girls and nine of them are boys. The researcher paid attention to the issue that these students attended the interviews done by the researcher voluntarily.

3.2.2. Instructors

Ten instructors, who took part in the final speaking exam, participated in the study. They are all English instructors at Zirve University Prep School. As the researcher was trying to identify the problematic areas and looking for solutions for these problematic areas of the speaking component of the final exam for Level B, the instructors were selected according to the level they were assigned in the final exam and their willingness to participate to the study.

Among 10 participant instructors, four of them were native speakers of English, who were also teaching Speaking and Listening classes at Zirve University; and six of them were non-native speakers of English, who were teaching classes specifically at Level B. Their age ranged from 24 to 45, and their experience in teaching English ranged from 16 months to 13 years. All of them at least teach 20 hours of English in a week to three to four different classes. Four of the participant instructors are females and six of them are males. All participants experienced speaking exams before, as speaking exams are given eight times in an academic year.

The Listening and Speaking Coordinator of the English Language Program, who is also a native speaker of English, is interviewed in order to collect data for content validity of the exam as well.

3.3. Data Collection Instruments

The instruments employed in this research study were Level B students' exam scores, speaking exam task, speaking exam rubric, an interview with the Listening and Speaking Coordinator, 10 interviews with instructors and 15 interviews with Level B students.

3.3.1. Students' Scores

Speaking component of the final exam constitutes 15% of the total final exam score. Students are graded and given scores out of 100, and later on 15% of this score is calculated in order to obtain the speaking score, specifically. Three instructors grade a student during a speaking exam. The interlocutor gives a score out of five and the graders give a score out of 45. The total speaking exam score of a student is calculated by taking the average of graders' score and adding the interlocutor's score.

2012-2013 academic year, first term, first quarter final speaking exam scores of Level B students were used for the quantitative part of the study. The reason of choosing level B students specifically was that Level B has the biggest number of students and this will provide a better amount of data. 750 Level B students' speaking exam scores were obtained and used in order to calculate the inter-rater reliability of the exam.

3.3.2. Speaking Exam Task and Speaking Exam Rubric

In order to find the answer for Research Question 1 (*How valid is the speaking component of the final exam at Zirve University Prep School?)*, the researcher did a content analysis by using the guidelines for the final speaking exam (see Appendix A), the final speaking exam task (see Appendix B) and the speaking exam rubrics (see Appendices C-D) which were used in this exam.

Speaking exam task and the rubric have regularly been changed so far. Since the final speaking exam, which was held on November, 27-28, was the issue of this research, only the tasks and the rubric used for this exam were taken into consideration.

For the final speaking exam the students are examined in pairs by three staff. One of the instructors acts as an interlocutor and the two other instructors are graders. The interlocutor direct the test and moderate if needed. The examiners do not speak to the students or with each other during the exam and grade individually. The graders assign a grade (out of 45 points) using task specific rubrics (see Appendix D). The interlocutor assigns a grade (out of five points) using a holistic scale (see Appendix C). The total score for the exam is graded out of 50 points. The test takes about 10 minutes per pair of students. The test is constructed of three tasks, each designed to target a different speaking skill.

Task 1: This task begins with introductions using a prompt sheet. The interlocutor then asks at least two follow-up questions to check for understanding. The students are also asked to spell something simple (e.g. surname) and give an extended number. The students can only interact with the interlocutor for this task. It takes about three minutes.

Task 2: For this task, students are required to ask for and give basic information based on very familiar situations. There are two situations. The students are given separate prompt sheets and then asked to formulate questions in order to get specific information. Their partner then answer the questions. This process is repeated for the second situations and takes about four minutes.

Task 3: This task is a consensus finding/discussion task. The students have one prompt sheet to share. They need to express their own opinions, discuss together and agree upon a given task. The students speak with one another for this task. It takes about three minutes. There are two kinds of final speaking exam rubrics. The interlocutors use a holistic rubric and give a score out of five. This rubric has five simple descriptors of required performance. The other kind of rubric is designed for the graders and includes three different parts which are structured for the three tasks in the exam. The points each task has are decided according to the difficulty of the tasks. Therefore, task one is graded with a score out of 10. Task two has a weight of 15 points and task three is graded with a score out of 20. While grading the graders first need to decide to which category the performance of the student belongs, and choose one of the steps named like "yes" , "partly" or "no" for each task. Then, as a second step, the graders need to choose the score for that performance. Each grader gives a score out of 45.

3.3.3. Interviews

Two different interviews were done for this research study. The first kind of interview was conducted with the instructors and the students, in order to answer the Research Questions 3 (*What do the instructors think about speaking assessment process?*), 4 (*What do the students think about speaking assessment process*?) and 5 (*What are the suggestions of instructors and students for a better speaking assessment procedure?*). This interview contains two parts. Part one contains four questions which are prepared to investigate the general features of the participants. Part two contains 12 questions which are prepared with an intent to investigate the thoughts and suggestions of the participants about certain aspects of the final speaking exam (see Appendix F). Ten interviews were conducted with instructors who took part in the final speaking exam. Moreover, 15 interviews were held with students from level B, as well. All interviews were audio-recorded, to be resolved and analyzed by the researcher later on.

The second interview (taken from Sak, 2008), which was prepared for investigating research question 1 (*How valid is the speaking component of the final exam at Zirve University Prep School?*), is for examining the content validity of the exam. This was conducted with the Speaking and Listening coordinator of the English Preparatory Program and contains 10 questions about final speaking exam (see Appendix E).

3.4. Data Analysis

3.4.1. Validity

Content validity of the exam was examined with the help of the analysis of the speaking exam task and the rubric and also from the data which was collected through the interview done with the speaking and listening coordinator. The interview includes questions about the content of the exam and the content of the speaking syllabus, and intends to find out how adequate the speaking exam is, in terms of containing relevant materials or tasks about the speaking curriculum. The interview was analyzed by the researcher, together with the speaking exam task and speaking exam rubric, in order to investigate the content validity of the exam.

3.4.2. Reliability

As Underhill states (1987, cited in Sak, 2008, p.48) "the classical measures of test reliability have little relevance for oral tests because they are designed for rigid, pre-planned tests. More useful information could be gathered by comparing each marker's scores with the scores of other markers." Based on this, the inter-rater reliability of the exam was estimated. In order to calculate inter-reliability of the exam, Pearson correlation coefficients of the 1st gradings of each pair were computed. That is to say, the scores assigned by each grader was correlated with their partners to identify how consistent the scores were.

3.4.3. Content Analysis

Content analysis of the interviews held with instructors and students, is done by the researcher in order to find answers to the research questions 3 *(What do the instructors think about speaking assessment process?*), 4 (*What do the students think about speaking assessment process?*) and 5 (*What are the suggestions of instructors and students for a better speaking assessment procedure?*). The study aims both to identify the strengths and weaknesses of the final speaking exam, and to find out the suggestions of both teachers and the students in order to have a better speaking exam procedure. Therefore, the interviews were analyzed by the researcher to examine the problems, solutions for these problems and new opinions and suggestions of the participants of the study by focusing on the commonly given answers.

3.5. Data Analysis Procedure

The data analysis was performed in five steps. Firstly, in order to find out the content validity of the exam, the interview about content validity was analyzed together with the speaking exam task and the rubric.

Secondly, level B students' speaking scores were used to calculate the reliability of the speaking exam by correlating the grades of each grader with their partners'. Thirdly, not only problematic areas of the final speaking exam, but also the suggestions for a better speaking exam procedure were identified through interviews held with the instructors. The researcher did this by analyzing commonly given answers of the participant instructors. The same process was done with the students as well, as a fourth step.

CHAPTER 4

4. FINDINGS

This chapter presents the data analysis and interpretation of the results. Firstly, content validity of the exam is analyzed and presented. Then, the results of inter-rater reliability analysis of the exam is given. The responses to the interviews, both from teachers and students are presented and discussed as the last step.

4.1. Analysis of the Data

This study aims to investigate the process of the speaking component of the final exam held in preparatory classes of Zirve University. Thus, data was collected from the interviews, content validity analysis and inter-rater reliability analysis of the exam in order to achieve this aim.

4.2. Validity Analysis

4.2.1. Analysis of the Responses to Content Validity Interview

Skill coordinators are the people who are familiar with the constructs and details of both that skill and that the exam measures. Therefore, in order to analyze the content validity of the final speaking exam, an interview was conducted with the Listening and Speaking skills coordinator of Zirve University English Preparatory Program, who prepared most of the exam. The analysis is done for each question one by one.

4.2.1.1. Question 1: Who is the test designed for? What is it designed for?

The responses of the participant reveal that the speaking test is designed for the students of Zirve University Preparatory Program for all levels. There are 4 levels, consisting of two months, in the program. Thus, final speaking exam is an achievement test done in the end of each term. As the one prepared for level B is the case for this research study, more specifically, it is designed for the level B students completing 108 hours of English education which contains 49 hours of speaking and listening classes. Levels in the whole speaking exam are supposed to be attached to the Common European Framework (CEFR). A level at Zirve University is regarded as the same level of A1 in CEFR. Other levels should be listed as this: Level B equals to A2, Level C equals to B1 and level D equals to B2.

For the second part of the first question, the participant states that it is designed to test the students' speaking skills; not the content, but the skill sets they covered in the classes. As this exam is a component of the final exam, and all other skills are also tested, this exam is mainly designed to test and assess the oral proficiencies of the level B students.

4.2.1.2. Question 2: What is the basis for considering whether the test is appropriate to your students?

The interviewee accepts that they do not have any certain basis, unfortunately. Speaking, as a skill to be developed, is considered as one of the important aspects at Zirve University Preparatory Program. However, there's a big gap between what the students are supposed to be and what they are. The reason for this, as stated by the interviewee, is that neither they have any standards or aims, nor a curriculum to meet some standards. The interviewee also mentioned that the speaking lessons are based on textbooks which are not even related to what it is supposed to be tested. Therefore, they end up with this huge gap in between what the students can do and what they are supposed to do. The example that the interviewee gives makes the situation clear that, level D speaking textbook is only an A2 level textbook; however, they are tested as B2 level students.

4.2.1.3. Question 3: Do you have any test specifications?

The interviewee states that technically, there are no test specifications developed for the exam; however, it is constructed by taking the CEFR descriptors. Regarding this, there are a number of things that paid attention while developing the speaking test. Firstly, the test includes the skill sets for really simple social exchange, which is also one of the things that done in the speaking and listening classes. Secondly, it is constructed to include a general information gap, like asking and answering some questions, which is the second skill set that emphasized in the speaking and listening classes. Third one, and most probably the most difficult one for the students is the discussion task. The interviewee mentions that these are the very specific tasks that they want to target in the speaking exams.

4.2.1.4. Question 4: Is speaking test content relevant to test specifications?

It's concluded from the answer of the interviewee that as they do not have any certain test specifications, it can be said that first they decide on the content of the test and then the test specifications shape. The interviewee clearly accepts that this is a very big problem and occurs because of the fact that there is no certain speaking curriculum to be taken into consideration while deciding on the content of the exam at Zirve University. Thus, it can be concluded that since there are no criteria and standards coming from the top, there is a bottom-top system while constructing the speaking exams.

4.2.1.5. Question 5: Do the items or tasks in the test match what the test as a whole is supposed to assess?

It's concluded from the answer of the interviewee that there is a parallelism between the tasks that the exam includes and the aim of the test. She states that the goal of a speaking exam should be to assess the speaking ability; and ability means not the content, but the skill. Therefore, the tasks used in the exam match what is supposed to be tested. She restates that there is a relationship between the tasks used in the exam and the aims of the exam as they are determined according to the CEFR descriptors.

4.2.1.6. Question 6: Does the test produce a good sample of the contents of the syllabus of the preparatory class?

The interviewee's answer to this question is a clear no. She mentions that the syllabus is written out from the textbook as there is no curriculum. Additionally, the textbook is a listening heavy textbook which means the syllabus becomes a listening heavy syllabus, as well and does not have emphasis on speaking. The only emphasis done on speaking is the creativity that the teachers put into it, and beyond any doubt, this changes from one teacher to another. To conclude, the speaking exam content is accepted by the interviewee to be completely different from the content of the speaking syllabus. However, this problem is tried to be solved by the skill coordinator by providing teachers with sample speaking exams to be studied and practiced in their speaking and listening classes.

4.2.1.7. Question 7: How well do tasks/ items of the test reflect the characteristics of speaking ability?

The interviewee states that the tasks used in the speaking exam work well in terms of reflecting the characteristics of speaking ability, they can not be named as perfect, though. They can also be accepted as authentic materials since real pictures are used. The exam also consists of several speech acts like greeting, describing, expressing ideas, discussing, agreeing and disagreeing. Thus, the exam reflects the characteristics of speaking ability well enough.

4.2.1.8. Question 8: What research was conducted to determine desired test content?

The interviewee admits that no research was done in order to determine the content of the exam as a whole institution. The exam is constructed only with personal research and efforts. She states that she examined the CEFR and its criteria in order to develop the speaking exam and modify those items as to meet the needs of the institution.

4.2.1.9. Question 9: What research was conducted to evaluate test content?

Similar to the question 8, the interviewee states that no research is done in order to evaluate the exam content. She states that preparing the exam takes so much time that when it comes to evaluate the content of it, almost nothing is done. Additionally, she mentions that the suggestions, complaints and feedback from both the teachers and the students are always considered.

4.2.1.10. Question 10: Are the tasks and topical contents relevant to the target language use domain namely, the potential uses, or the situations that the test taker is likely to encounter?

In order to answer the question, the interviewee examines the speaking exam tasks and states that they can be regarded as relevant to the target language use domain. She clarifies that the aim while preparing the exam is to test very basic and required things that the students should know and also ask the teachers to fit their teaching styles to these. Therefore, there should be consistency between the tasks of the exam and the language domains. She summarizes that the first part includes greetings and basic questions about the students. The second task is an information gap activity which can regarded as one of the situations that all the students are likely to encounter. The last task is a discussion task, which requires students to display their analytic thinking skills and state agreeing and disagreeing on o topic. This task can also be considered as relevant since what the students need in their departments and their future lives is analytic thinking, judging, making conclusions and expressing their ideas.

Having all the answers to the questions about the content validity of the final speaking exam, it can be concluded that the interviewee is of the opinion that the exam can not be said to have content validity. It is also seen that the exam is not directly based on the content and the objectives of the speaking and listening classes. Course content is not used to construct the speaking exam; on the contrary, speaking exam is the element that is used to decide the course content.

4.3. Reliability Analysis

Since the final speaking exam is a kind of test where raters' personal opinions and judgements are of big importance and affect the scores for the performances, inter-rater reliability level of the exam is investigated. The results are displayed in tables.

4.3.1. Inter-rater Reliability

In order to calculate inter-reliability of the exam, Pearson correlation coefficients of the first gradings of each pair were computed. That is to say, the scores assigned by each grader was correlated with their partners to identify how consistent the scores were.

Table 1. *The Correlation Coefficients of Each Pair's Ratings*

		Rater 1	**Rater 2**
Rater 1	Pearson Correlation	1	,941**
	Sig. (2-tailed)		,000
	N	750	750
Rater 2	Pearson Correlation	,941**	1
	Sig. (2-tailed)	,000	
	N	750	750

In Table 1 the correlation coefficients displaying the inter-rater reliabilities of the two graders are given. The scores from each grader were lined up in columns and the results were gained by calculating the correlations between these two scores on SPSS. As shown in the table, the correlation coefficients from the two graders are 0.941, indicating fairly high inter-rater reliability. It is concluded from the table that the raters agree with each other while rating.

4.4. Analysis of the Responses to the Interview Held with the Instructors

The interview, which was held with the instructors, aims to find out the thoughts, complaints and suggestions of the instructors who participated in the exam. The interviews, which contained 11 questions, were all audio recorded and analyzed by the researcher according to the common answers. 10 instructors participated in the research and volunteered for the interviews. 4 of them were assigned as interlocutors in the final speaking exam and were native speakers of English. The rest 6 were Turkish teachers who were the graders in the final speaking exam. All answers from each teacher are collected and content analysis is done for each question one by one according to the commonly given answers.

4.4.1. Question 1: What do you think about the final speaking exam procedure in general?

The aim of this question is to find out the participants' general ideas and feelings about the final speaking exam. According to the results, the interlocutors generally accepted the exam as good enough. It is said that this type of exam is definitely a step to the right direction as it is the first time that the ability of communication is tested rather than just asking questions to find out if the students learned the materials in the book or not. One of the interlocutors criticizes that some tasks are unnecessary and the exam should only assess fluency, nothing else, it does not, though.

Graders' general feelings for the procedure of the exam are also good. They mostly answered positively. It is accepted that this exam is better than the previous ones, as it enables students to have a real-like communication instead of just having students ask and answer some questions. However, it is also stated that this new type of speaking exam procedure is still too complex and difficult for the students as it is something new. One of the graders said that the students generally were not as active as he had expected, and he believed the reason of this was the procedure of the exam. Here are some excerpts from the answers:

Interlocutor 1: "I liked it. It is a right way to assess speaking ability."

Interlocutor 2: "I think it is good. There are some points need to be changed, though."

Grader 1: "It was OK. I like this way of testing speaking."

Grader 2: "I believe it is a little bit difficult for the students. However, it gives the students the chance to have a real communication."

Grader 3: "It is better than the previous ones."

4.4.2. Question 2: What do you think about the “time” issue in the final speaking exam? Is the amount of time allocated for each student enough?

In order to find out the participants’ thoughts about the duration of the exam this question is asked. From the interlocutors point of view, the exam is too long. It is more than enough both for the students to express themselves and for the teachers to evaluate the students’ speaking ability. The process of the exam is defined as so long that it becomes something exhausting and tiring. One of the interlocutors stated that he could have given the students a mark even after the second task of the exam, there was no need for the third task, which was the most difficult.

For graders, the exam is judged as too long, as well. They say it is much more difficult for the graders that all they do is to sit and listen to the performances for hours and grade them. This is much more tiring as the interlocutors have the change to speak to the students. In addition to these, all graders accept that the students who were the very first ones on the list were luckier because graders admitted that they could not pay the required attention to the students who take the test later. As it is a quite long exam and this makes the graders too tired to listen to the performances, assessing becomes something certainly unreliable. On the contrary, only one grader thinks that timing is good since students are graded in pairs. The grader states that trying to grade the students one by one is a slower and more tiring process than grading them in pairs. Here are some examples from the answers:

Interlocutor 1: “It is more than enough. This makes the exam tiring both for the teachers and the students.”

Interlocutor 2: “We did two sessions of test which were exhausting and too long.”

Grader 1: “It was very difficult for me to listen to all those performances for hours and try to grade them all.”

Grader 2: “I think it is very very long. I could have graded the students in shorter periods.”

Grader 3: “It was long but having students one by one would be longer. Pairing made it shorter.”

Grader 4: “I have to admit that I could not pay enough attention to the performances after some time, as the exam was so long.”

4.4.3. Question 3: Which one do you think is better for a speaking exam; having the students one by one, in pairs or in groups? What are your reasons?

There are many kinds of doing a speaking exam and the number of the students taking the exam at the same time is an in important factor. To find out the better ways for the participants, this question is asked. Interlocutors generally answered the same way and said that pairing was the best way. Grouping, which is tried before as a speaking exam at Zirve University, is accepted that too difficult to control and follow what's going on. In addition to this, some students may be lost in groups and stronger ones can take each role. Thus, this makes the exam for the interlocutors too difficult to control and chaotic both for grades and interlocutors. However, pairing is accepted by the interlocutors to be the best way, as it makes the students talk and force them to discuss. They all agree that taking the students one by one is not a good way because then, students just give a short answer and wait. It's more difficult to help them speak more. One of the interlocutors criticized that the exam format was wrong because the same format and the same task types were used for all levels. He said for levels A and B the exam should have been done individually, but for higher levels, C and D, it could have been done in pairs.

Graders also seem to agree on nearly the same ideas. They say that having the students one by one is the most stressful way of doing a speaking exam for the students. Additionally, it is not good as the students only speak to the teacher, which will not be a real conversation and make them more nervous. Pairing, on the other hand, is a better way both in terms of time and grading, as it's more interactive and real-life like. Graders also state that this is a way that reduces students stress as they are with their friends and not alone in front of a teacher who is asking questions. Still, one negative aspect of pairing is also stated that if the pairs are not chosen carefully, they may affect each other negatively which results in a bad performance. Lastly, same with the interlocutors, grouping is the way that is not liked by the graders as well; since it is very difficult to control and grade. It is repeated that some students say nearly nothing and be lost in the group as the others have the control of the group, and there is nothing to grade for the grades. Here are some excerpts from the answers:

Interlocutor 1: "I think pairs is the best. However, level A students may have difficulty with this format. They should have taken the exam one by one. Each level should be different."

Interlocutor 2: "Grouping would be very difficult to control, I do not like it. Paring is OK. It makes them speak with their partners "

Interlocutor 3: “I believe individually done exams lead students memorization. He only makes a sentence and waits for me to help him.”

Grader 1: “Being individual is not good as students only speak to the teacher. That’s why pairing is a better way.”

Grader 2: “ Pairing is good but pairs should be chosen very carefully. A student can affect the other easily. ”

Grader 3: “Being with a friend in an exam is less stressful. I like pairing.”

4.4.4. Question 4: What do you think about the format of the speaking exam? Which one would be better; interviews, presentations, discussions and so on? Do you have any suggestions for the format of the exam?

Speaking exam formats are another issue to be investigated and analyzed. According to the analysis of this question, it is found out that interlocutors all agree that doing presentations is the worst way of doing a speaking exam since it leads students to memorize. In this exam pair discussions is used and it is said to be very good; however, interlocutors also state that the students need to do more practice with this type of format before the exam because it seems doing discussions is very difficult for them to handle. Especially for A and B level students the exam is very difficult as doing a discussion requires analytical thinking.

Graders generally agree on the idea that presentations should not be used as the students all memorize, so the grading can not be fair. Interview, which is too restricting as it only requires question-answer tasks, is not a good way either. Discussions seem to be a better way, our students need more practice, though. Only one grader stated that presentations, in which the students use some multi-media components and by this way, express themselves freely, would be a good way too. Two of the grades agree on the idea that, the more different format types a speaking exam has, the better the exam will be. Therefore, they suggest to use all kinds of format types as much as possible. Some excerpts from the answers are like these:

Interlocutor 1: “It is generally OK. Discussion is difficult for lower levels.”

Interlocutor 2: “We should not use presentations as students only memorize.”

Interlocutor 3: “Using discussions in all levels is not correct. It is difficult for low levels.”

Grader 1: “Discussions are he best but our students are not ready for this.”

Grader 2: “The ideal way would include as many different tasks as possible.”

Grader 3: “ Presentations are not good. However, as they are able to use multi-media they will feel more relaxed with their presentations.”

4.4.5. Question 5: What do you think about the importance of using visuals aids in the speaking exam?

Using some aids, especially pictures has a notable affect while having an exam. The participants’ ideas and beliefs are questioned with this question. Interlocutors believe that using pictures in the exam helps a lot. As they are still learners, giving them something to see, like a trigger in their hands, is definitely very useful and helpful. The students are always nervous in the exams, so giving them something to talk about and use is needed. Although the idea of using visual aids is accepted, one of the interlocutors thinks that the pictures used in the exam can be chosen better.

From the graders point of view, using visual aids in the speaking exam is considered highly significant. These can be concluded from the grades answers that, pictures are good prompters when students stop and allow students to have a initiation to start their speaking with. Visual aids also help brainstorming, which eases the speaking performance of the students and gives the students a starting point when they do not understand the question completely. Here are some of the answers given to this question:

Interlocutor 1: “It is very useful to use pictures. It helps a lot.”

Interlocutor 2: “Pictures are like starting points fro them. I also believe some pictures in the exam should be changed.”

Grader 1: “It gives them clues to speak and helps brainstorming.”

Grader 2: “They are very useful as they give students the chance to start easily.”

4.4.6. Question 6: What do you think about the grading procedure of the speaking exam?

The most important part of a speaking exam is the grading procedure. Grading tools, rubrics can change a good exam into a disaster. Therefore, the quality of the rubrics used in final speaking exam is analyzed by investigating the participants ideas. One of the interlocutors said that he did not like the holistic rubric and found it distracting to try to talk and to think to grade at the same time. Others think the holistic rubric is good and clear enough. They say that they are, as interlocutors, happy with using such kind of a holistic rubric and give a score out of 5. If they gave more than that percentage, as they were the ones who had to lead the exam, it would not be a fair grading. One of the graders criticized some items on the rubric that the grades use, and say that there should not be any criteria for grammar and vocabulary on a speaking exam rubric, since we should be assessing only the ability to speak, and the fluency in a speaking exam.

When we examine the grades' answers, this can be easily understood that almost all of them do not like the rubric that they are using. Even one of the graders states that grading part is the most problematic part of the exam and say the rubric is not practical at all. They also make the point clear that, trying to listen to and grade too many students in a day is very tiring and distracting that the grading turns into something certainly unreliable. If the teachers do not like the rubric or are not explained how to use the rubric affectively, they admit that they do not follow it and just give standard, nearly the same grades to all students. Furthermore, if the rubric is not structured well enough, graders‘ personal features like being positive, helpful or strict can not be eliminated, and this makes the grading unfair and unreliable. However, they accept the idea that one student is graded by 3 different teachers is something good. Here are some excerpts from the answers:

Interlocutor 1: “I think it is easy to use these rubrics. Giving a score out of 5 is good, otherwise it would be more difficult to grade fairly.”

Interlocutor 2: “Holistic rubric is OK but the one that graders are using should be changed. Grammar and vocabulary should not be tested in a speaking exam. We have separate exams to test these.”

Grader 1: “I think having these rubrics is the worst part of the exam. I do not like them.”

Grader 2: “Rubrics should be considered again carefully. Having three teachers grading the performance is very good. Despite, we can not stop the teacher's personal features to take part in grading as well.”

4.4.7. Question 7: What do you think about the physical circumstances (exam layout, seating etc.) during the speaking exam?

Another issue to be investigated is the physical circumstances of the exam. In order to examine this aspect about the exam the participants' ideas are sought. The interlocutors like the seating and the idea behind it and admit that it is better for the students not to see the graders. Only one of the interlocutors states that he does not believe it will change anything for the students to see the graders while they are grading as the students are always nervous in the exam.

Graders' answers to this question vary. Some of them believe that it is correct that the graders sit behind the students, so that students do not see them while grading. They say that this helps the students to feel less nervous. Others state that to see the gestures and faces of the students would make the grading easier and more fair. Without seeing these, they admit that they have difficulty to pay attention to the performances after some time. Two of the graders also mention the issue that there are too many students waiting outside and the noise coming from them is problematic both for the students who take the exam at that time and the teachers who are grading. Some examples of the answers are like:

Interlocutor 1: "The seating and the arrangement of the chairs were OK."

Interlocutor 2: "I believe it is correct that the students should not see the graders during the exam."

Grader 1: "The students only interact with the interlocutors, which helps the students not to get stressed. "

Grader 2: "I feel like I have to see the students' faces and gestures while grading. Otherwise I can not concentrated on their performances."

Grader 3: "The students outside of the class make too much noise which I found irritating."

4.4.8. Question 8: Do you think it is better to assign only native speakers as interlocutors and Turkish staff as graders? What do you think about examining your own speaking-listening class?

Being native or not is always accepted one of the most important factors about assessing speaking. That is why participants' ideas about this issue are investigated. All the interlocutors' answers show us the same thing that they believe having Turkish staff or native speakers of English as interlocutors does not change anything. They say that it's better for students to get used to different accents of English. Moreover, they also state that if the students have a chance to see their own listening and speaking teachers as interlocutors, this will be the best and the most comfortable. As the students know their teachers and their accents, it helps a lot, and they will be less nervous. On the other hand, they also expressed that if the graders were the native speakers, then this would affect the scores noticeably. Their reason for this is that, Turkish teachers generally have a tendency to focus on grammar and vocabulary, rather than fluency, pronunciation and comprehensible speech. They conclude that there can be two different options; training the Turkish graders in a real sense, or having one Turkish teacher and one native speaker as graders.

From the graders' point of view, the answers are different. Two of the graders state that international staff should be assigned as interlocutors, not the Turkish teachers because it sounds more natural when they speak and a speaking exam should be done like this. However, most of the other graders claim that there would be no difference with the Turkish teachers or native speakers as interlocutors, since being native or not has nothing to do with grading or doing the exam better. In addition to this, they also believe that the students should not have their own listening and speaking teachers as interlocutors because in such a situation, the interlocutor can not be fair enough as he knows his students, can also affect the graders, and this results in a grading which is not objective. Some also claim that Turkish staff at Zirve University may be better interlocutors as they believe some native speakers come from different backgrounds and do not have the required qualifications of a teacher. Some examples of the given answers are:

Interlocutor 1: "It will not be a problem to have Turkish staff as interlocutors, accent is not everything."

Interlocutor 2: "Turkish teachers mostly focus on grammar so it changes a lot about grading."

Interlocutor 3: "I think it is not about being Turkish or a native. However, students get used to their own speaking and listening course teacher, so using their own teachers would be the best."

Grader 1: "We can not say that native teachers are better interlocutors since they are native. I think it would change nothing."

Grader 2: "I believe having native staff as interlocutors is correct."

Grader 3: "Listening and speaking course teachers can not be objective about their own students, so it should not be tried."

4.4.9. Question 9: Do you pay attention to include some aspects of delivery such as body language, eye contact and so on into the assessment criteria for speaking?

It is accepted that as assessing speaking ability is something subjective, consciously or unconsciously, assessors may sometimes pay attention to some other aspects of the performance. In order to find out the situation in this research case, it is asked to the participants. For the interlocutors, such kind of issues should just have a role in order to help the teachers to understand the students' psychological situations during the exam, but should not affect the grading. They all say that they try not to pay attention to these while assessing a student.

Graders have completely different ideas about this issue. They believe that although students' body language is said not to be assessed, it should be; since using such kind of things correctly shows the real ability to speak and the mastery of using the language. They also state that communication is a whole and it includes body language as well. Moreover, they add that especially teachers should use their body language and gestures affectively, consolidating the questions with body language definitely helps students to understand better. Only one of the graders claims that it should not be criteria for assessing as the students are very nervous during the exam. Below are some examples from given answers:

Interlocutor 1: "These help me to realize their mood, I do not pay attention to them while grading."

Interlocutor 2: "I only try to pay attention to their speech, not anything else."

Grader 1: "Body language of the students is important as it shows how professional they are about using that language."

Grader 2: "Communication includes facial expressions, gestures and body language as well, so I pay attention to them."

Grader 3: "I think we should not grade according to these aspects because students are always very nervous during the exam."

4.4.10. Question 10: What is the aspect of the final speaking exam you liked most?

In order to find out the strong features of the final speaking exam, answers to this question is examined. Interlocutors mostly like the format of the test that including discussions, which leads students to try to speak instead of giving short answers. They also agree on the idea that this is a better way to assess speaking and defend the whole idea of assessing "Do the students actually speak?", instead of testing information. Another aspect which is regarded as good by the interlocutors, is that the exam has three different tasks, from the easiest to the most difficult. The first task is the easiest which helps the students relaxed and get into the speaking mode.

The thing that made the graders happy about the exam is to see that students really tried to speak and some of them did this very well. This type of exam helps students to show their ability to communicate in a real sense and noticing their success is a good thing for the graders. Furthermore, in previous exams students can easily learn the questions from the ones who has just taken the exam before them, and memorize the same answers. This time they do not know the content of the exam, which is considered another good aspect of the exam by the graders. Here are some excerpts from the answers:

Interlocutor 1: "The tasks were from the simplest one to the hardest, I liked that. This helped them to have a better start and be motivated for the rest. "

Interlocutor 2: "I liked the format of using discussions. By this way we could really able to assess their ability to speak."

Grader 1: "I noticed that my students tried hard to get to speak. I felt happy about that."

Grader 2: "Students did not know the questions and the content of the exam. I think this is the best thing about the exam."

Grader 3: "With the help of pairing and discussions memorization was eliminated. This can be considered as a good aspect I believe."

4.4.11. Question 11: What is the aspect of the final speaking exam you disliked most?

In order to investigate the weakest points about the exam generally the participants' answers to this question are analyzed. The aspects which are mostly disliked about the exam by the interlocutors is that it is too long and tiring. Two of them state that they could have assessed the students with the very first tasks and that there is no need for the last task.

Graders agree with the interlocutors on the issue that the duration of the exam should be shorter as it is too exhausting for the graders to listen and grade all those students in a day. Another thing which is not liked by the graders is the rubric; they believe it should be revised and tried before the exam. Lastly, one grader states it is not good that the interlocutors can not help the students who stick and can not go on the speech as it's forbidden to do paraphrasing and change the script. Some examples can be stated like:

Interlocutor 1: “The duration. It was very tiring.”

Interlocutor 2: “Time was too long. I did not need that much time for grading.”

Grader 1: “It was exhausting because the tasks were very long.”

Grader 2: “I really do not like the rubric at all.”

Grader 3: “Maybe the interlocutors can have the right to help students to some extend. I do not like the situations that the students get stuck because of only a word.”

4.4.12. Question 12 : How could the test be improved? Do you have any suggestions on how to improve the procedure of assessment of speaking?

In order to suggest better ways to carry out the exam and the things suggested to be changed, participants’ answers to the question are analyzed. There are many suggestions about the final speaking exam from the interlocutors. They state that a leader of the test is needed to be decided. It should be clarified who is responsible and in charge of the all paper work, entering the grades and submitting them and also the videos to the testing office. This person is suggested to be the interlocutor of the exam. By this way the procedure from the beginning to the end of the exam becomes clearer.

Another suggestion from the interlocutors is that the rubric that the graders are using should be revised and restructured. Rubrics should not be complicated and the language used both in the holistic rubric and the graders’ rubric should be simplified. Also the items about grammar and vocabulary should be changed or removed completely.

In addition to these, they suggest to change the discussion task. It is said to be a better assessment tool if there are not any restrictive questions like “where?”, “When?”, “What to do?”. These questions make the students speak in the same way; without them they will have more to discuss and speak freely. Discussion task in low levels (A and B) is needed to be removed or simplified as it requires analytic thinking which is rather difficult for those level students. Therefore, the format of the exam is suggested to be changed according to the levels. Each level should not include the same format of the exam.

Last but not least, they mention that it’s good to have clear instructions and a script; however, some of the interlocutors are known to change the script of the exam which, makes the exam unfair and subjective. Thus, there should be more training both for the graders and the interlocutors. They need to be explained what and how to do is the correct way for a speaking exam.

Finally it’s stated that both the procedure and the materials in the exam is relatively different from what is being taught in speaking and listening classes; however, the main problem is not this, but the speaking and listening curriculum. They all mention that Zirve University English Preparatory Program still has not got a real speaking and listening curriculum and there needed to be one urgently. They accept that, as speaking and listening teachers, all they have is the list of the contents page of the book that they are using, and they believe this results in a gap between what is being taught in the class and what is being assessed in the exam.

From the graders’ perspective there are some similar suggestions. The very first suggestion is that grader and interlocutor training is of importance and should be done regularly. All the teachers should be explained clearly that not the accuracy, but the fluency is the point to be the main focus in a speaking exam.

They accept that this type of exam is the one needed; however, we should have chosen different books in order to make this exam more affective. The students are needed to study critical thinking skills more often as there is a big difference between what is expected from the students in the exam and the content of the books they have been studying in the class so far.

Another important suggestion is that pairing is good, yet the pairs should not be chosen randomly as they affect each other easily. The listening and speaking teachers should be careful while deciding on the pairs; weak students should be paired with stronger ones.

Lastly, some of the graders think that the students need more freedom and will perform better in case they have it. For example, they should have been given quite a lot more pictures and the chance to choose the picture they want to speak about. Furthermore, interlocutors should be permitted to help the students who have difficulties while speaking. Some excerpts from the answers given are like:

Interlocutor 1: “I think it would be better if the rubrics are simplified. Maybe the graers’ rubric can be restructured.”

Interlocutor 2: “I would suggest to make the discussion part more uncontrolled and remove the question words in the tasks.”

Interlocutor 3: “More training sessions should be done I believe. Moreover, the curriculum problem should be solved in order to have more valid and fair speaking exams.”

Grader 1: “Listening and speaking teachers should be more careful while deciding on the pairing lists.”

Grader 2: “I think the rubrics should be changed or reconsidered about.”

Grader 3: “Teachers should have more training before the exam and be reminded that fluency and comprehension are main issues about grading a performance.”

Grader 4: “I suggest that we should let the students decide what to speak about and the interlocutors help the students when they need. ”

4.5. Analysis of the Responses to the Interview held with the Students

The interview, which was held with the students, contains the same questions as in the one done with the teachers, and aims to find out the thoughts, complaints and suggestions of the students who participated in the exam. Only the question 9 is changed like: “Are you affected by the body language of the interlocutor? How?”. The interviews, which contained 11 questions, analyzed by the researcher according to the common answers. 15 level B students participated in the research and volunteered for the interviews. All answers from each students are collected and content analysis is done for each question one by one according to the commonly given answers.

4.5.1. Question 1: What do you think about the procedure of the final speaking exam procedure in general?

The aim of the very first question is to have the students tell how they feel and what they think about the exam in general. The first thing that can be concluded from the students’ answers is that, they all think that it’s useless to assess the speaking ability as they can not show their knowledge and ability in an exam. They believe it should be done in another way, or should not be done at all. Some of them state that the reason for why it is useless is because they will not need to speak English in their departments, since education in some of the departments at Zirve University is 30% English, some others have 100% English education, though. All of the students agree on the idea that the exam is very difficult, especially the 3rd task, which is a discussion activity. They say 2nd task, which is a description activity is OK but for the 3rd task they mostly can not understand what and how to say. Even the topics in the discussion task are too unfamiliar for them and completely different from the ones they have been studying in the class. They all believe that the teachers are expecting too much from them, that’s why they seem unsuccessful. Only two of the students say that the exam is good enough to assess who is good and who is bad.

They think it is not easy because the questions are good and asked in the required way. Only one student admit that the exam was bad for him because he does not study at all. Some example answers are:

Student 1: “I am always very nervous in the speaking exam so it can not assess my ability to speak. I do not like having speaking exams.”

Student 2: “I will not need it in the future. Why do we have to have speaking exams? I think they are useless.”

Student 3: “The exam was very difficult. I did not know anything about those topics.”

Student 4: “I do not like the tasks. I think the exam is very difficult for me. I do not like studying for speaking.”

Student 5: “If you study hard you can do well in the exam. I believe it is a good exam.”

4.5.2. Question 2: What do you think about the “time” issue in the final speaking exam? Is the amount of time allocated for each student enough?

Duration of the exam is also needed to be analyzed. Therefore, the participants answers to this question are investigated. All answers from the students link on the same point that the exam was too long. Most of them think that the time was so long that they can not think of anything more to say or add. They say as it is too long, they feel more nervous and stressed and this also affects their performance. They admit that even after they finish their answers, seeing the teacher is waiting for more, makes them feel that they are unsuccessful and will fail from the exam. Some students make the issue clear that, even if it’s OK for the ones who are taking the exam at that moment, the others are waiting outside for hours, which is definitely something tiring and more stressful. Only three of the students say that the duration of the exam is good and enough for them. some examples from the answers are:

Student 1: “When I finished everything to be said, the teacher was still waiting answers. I think time was too much for me.”

Student 2: “It was very long. I was too tired after the exam. I do not like this.”

Student 3: “I believe it was OK. It was enough for me to speak.”

4.5.3. Question 3: Which one do you think is better for a speaking exam; having the students one by one, in pairs or in groups? What are your reasons?

Maybe one of the most important aspect of a speaking exam for students is the way they take the test. In order to find out their ideas about this issue, the participants' answers are analyzed. We can conclude that half of the students thinks pairing is good for them, the other half believes that taking the exam individually is better. The main reason stated being for individual exams is better is that it is less stressing and that they are better when they are alone. They say that they do not like taking the responsibility of their pairs and this makes them more nervous. Another reason is that, they believe if their pairs are not able to speak well, this affects them and they can not speak as well as they normally do either. One student states that just in case of having the chance to decide their own pairs so that they can practice before the exam, it can be a good way. Otherwise, it is not a good way as pairs easily affect each other.

On the other hand, half of the students believe that pairing is a good way as being with their friends, instead of being alone, helps to reduce their stress during the exam. They also state that if one pair gets nervous and stops talking the other pair can go on and they can help each other's performances. Some excerpts from the answers are:

Student 1: "I may cause something bad during the exam and my pair will get a low grade. I think it is not good to be with your friends in the exam."

Student 2: "I feel more relaxed and good when I'm alone in front of my teacher."

Student 3: "I think being with your friends in the exam is very good. If I stop talking, my friend can help me about my conversation."

Student 4: "I do not like the idea of having my friends with me."

4.5.4. Question 4: What do you think about the format of the speaking exam? Which one would be better; interviews, presentations, discussions, role plays and so on? Do you have any suggestions for the format of the exam?

Having the students ideas about the format of the exam that they have taken is important, so their answers about this issue are analyzed. It can be concluded from the answers of the students that, contrary to the teachers', their favorite speaking exam format is doing presentations. The most important reason for this is that they believe, as they know the topic that they will present, they have the chance to study for the speaking exam beforehand; as a result, they will feel comfortable and get less stressed during the exam and get better grades. Only one student states that he believes doing presentations is not a good way of a speaking exam as they will just memorize and learn nothing. They all agree on the idea that discussions should not be used in speaking exams. Instead, they suggest interviews can also be used. Some example answers to the question are like:

Student 1: "I hate discussions. I think we should not use them. They are very difficult."

Student 2: "I believe presentations are the best because I would have a chance to study for my exam like the other exams. "

Student 3: "I like interviews and presentations. We should only have them."

4.5.5. Question 5: What do you think about the importance of using visuals aids in the speaking exam?

In order to find out whether or not using visual aids is useful from the students' point of view, their answers are analyzed. The students generally like the idea of using pictures in speaking exams. However, they all mention that they want to have the chance to choose the pictures that they will talk about. Some of them state that they feel like they are bordered with the same number of pictures and questions and that they do not want to talk about only two pictures that the teacher chooses. Actually they think that the pictures used in the final exam are irrelevant, so it is very difficult for them to speak about those pictures. Another thing that the students do not like about the pictures used in the final speaking exam is that the questions asked with the pictures, whose aim is to lead the students. They say they should speak freely about the pictures instead of trying to answer the questions. One student conclude that using pictures is not fair because if one student has a good imagination he can speak better and get a better grade with the help of the picture. Some excerpts from the answers are:

Student 1: "Pictures are OK but I do not like the questions like -where, when, how- asked with the pictures."

Student 2: "I would like to choose the picture that I will talk about. This way I can speak better."

Student 3: "I like using pictures because I feel I can speak well with the pictures."

Student 4: "I think the pictures in the exam are meaningless. I should have the chance to decide about the pictures."

Student 5: "There should also be some more topics because if there are only pictures, some students like me will have problems about talking about something visual. I sometimes can not find anything to comment on about the picture."

4.5.6. Question 6: What do you think about the grading procedure of the speaking exam?

As being graded by three different teachers, students' ideas and feelings are of importance. In order to find out their ideas about the grading procedure, question six is asked. The most common answer for this question is that the students believe that the percentage of the interlocutors' grade should be higher than the other two graders as interlocutor is the person whom they have interaction with. Even one of the students mentioned that she did not want the graders to grade her performance as she could not be sure if they were really listening to her performance or not. For some others this grading system is good as it would be unfair and very difficult for the interlocutor both do the exam and grade the students at the same time. One student say that it's a fair way that they have been graded by three different teachers. Another student suggests that grading should not be done during the exam and that teachers can come together after the exam and grade their performances according to the videos. By this way speaking exams will be less stressful. Here are some examples of their answers:

Student 1: "I do not believe that the graders always listen to my speech carefully. I would like to have only the interlocutor do the grading."

Student 2: "Three people are better than one person because by this way we have a chance to get a better grade from at least one of them."

Student 3: "I think it's OK. Otherwise the interlocutor can not do everything well."

Student 4: "Why don't they come together after the exam and do the grading later from the videos?"

4.5.7. Question 7: What do you think about the physical circumstances (exam layout, seating etc.) during the speaking exam?

Physical circumstances are other important factors that affect students' performances. That is why their ideas are sought about this issue. There are three different answers given to this question. The first common answer is that students think only one teacher should grade their performances. Knowing that there are two teachers sitting behind them and grading their performances, they get more nervous. Thus, they defend that only the interlocutor should give the grades. The other common answer is that some of them like the idea of having two teachers behind them as they will be stressed much more if they see them grading. The last answer is that the graders should sit in front of the students, not behind them, in order to see the students and understand their speech. By this way they can grade more fairly. Here are some of the excerpts from their answers:

Student 1: "I always feel stressed because of the teachers sitting behind me because I am aware that they are grading my performance."

Student 2: "I would not want to see people writing notes about my performance while I am talking. I believe teachers should stay behind me so that I can not see them."

Student 3: "I am not sure that the teachers sitting behind me listen to me and my performance. Maybe it would be better that I can see what they are doing during my speaking performance."

4.5.8. Question 8: Do you think it is better to assign only native speakers as interlocutors and Turkish staff as graders? What do you think about examining your own speaking-listening class?

Students' perceptions about native and non-native teachers have always been a significant issue in this area. The participants' answers are analyzed and it is found out that most of the students state that having their own listening and speaking teachers as the interlocutor in the speaking exam would be the best way to do. Their reasons for this are that their teacher knows them very well and that they will not be nervous with him/her. Another common answer is that they would like to have a Turkish teacher as the interlocutor in the exam, since they have difficulties understanding other native teachers. They even say that all of the teachers doing the speaking exam should be Turkish teachers. Two students mention that they would not like to have their own teacher as the interlocutor because they believe he/she can not be objective about his/her own students. Only one student accepts that having another native teacher, rather than their own teacher as the interlocutor, and Turkish teachers as the grades is the best way to the a fair and objective speaking test. It can also be concluded that some of the students believe that Turkish teachers do not give high grades to the speaking performances of the students, but native teachers would. Some of the examples of the answers given are:

Student 1: “I want my own teacher to do the exam and the grading. I would feel better.”

Student 2: “I do not understand native teachers. I would like to have a Turkish teacher as the interlocutor.”

Student 3: “I think Turkish teachers should do the exam so that they can help us.”

Student 4: “My speaking and listening teacher will not give a good grade to me because he knows me and that I’m not good at his class.”

Student 5: “I am sure that Turkish teachers would give lower grades, so they should not do the grading.”

4.5.9. Question 9: Are you affected by the body language of the interlocutor? How?

One of the most important things that affects the students’ performance is the interlocutors’ attitudes towards students. That is why the participants‘ ideas are asked about the issue. It is concluded from all of the answers that body language of the teacher in a speaking exam is crucial for the students and affects their performances highly. They all give nearly the same answers that they believe the interlocutor should be cheerful, relaxed, smiling, energetic and motivating. Some of the students even believe that the interlocutor’s body language and gestures are the reasons of their failure or success. They mention that if the interlocutor is nonreactive, angry or not smiling, they get more nervous and forget everything to say. Another important thing to be concluded from their answers is that they all want the interlocutor to look at them directly, nod and show that he is really listening to their performance. If the interlocutor takes notes or deals with the papers he has, this affects them and their performance negatively. Here are some example answers given to the question:

Student 1: “The teacher in the final exam was nonreactive, so I can not speak because of him. ”

Student 2: “I’m affected by the teacher’s behaviors and gestures. If he does not smile I can not speak.”

Student 3: “The teacher should look at me and not deal with anything else so that I can understand that he is listening to me.”

Student 4: “I believe the interlocutor’s behaviors affect at least 50% my performance.”

Student 5: “If the teacher is relaxed and smiles I feel relaxed. He should motivate me.”

4.5.10. Question 10: What is the aspect of the final speaking exam you liked most?

The strong points of the final speaking exam are tried to be found out by asking the participants' ideas. The most common answer to this question is that there is nothing to like about the final speaking exam. However, some of the students agree on the issue that the 1st task, which is the easiest task of the exam that includes simple introducing oneself questions, is the only thing that can be liked. Still, only one student adds that 2nd task, in which they describe the pictures that they are shown, is also something good. Some answers are like:

Student 1: "First task was easier. I liked that."

Student 2: "I do not like anything about the exam."

Student 3: "I think describing pictures is good. I could do that part."

Student 4: "I like the task we start with. Asking and answering questions is good and easy.

4.5.11. Question 11: What is the aspect of the final speaking exam you disliked most?

Students' ideas about the weak points of the exam are sought. Many issues can be listed as the common disliked aspects of the final exam by the students, like: having three teachers, which makes the exam more stressful, waiting outside too long, pairing, discussion, unfamiliar topics that have no relation to the ones studied in the class, difficulty of the exam, duration of the exam, unsmiling teachers, having only two pictures to speak about and not having the chance to choose both the pictures and the topics, taking the speaking exam in a separate day and being have to wait for it. Here are some answers from the participants:

Student 1: "I do not like having three different teachers at the same time. It made me nervous."

Student 2: "The duration of the exam is the thing I do not like. I hate waiting a day for speaking part of the final after having sit down sessions. I do not like waiting outside for my turn either."

Student 3: "I do not like the exam as a whole. I think it is very difficult for me."

Student 4: "I dislike having discussions. I am not good at it."

Student 5: "I think the pictures and topics are very difficult. I do not like them."

4.5.12. Question 12: How could the test be improved? Do you have any suggestions on how to improve the procedure of assessment of speaking?

Students' ideas and suggestions are sought for determining what needs to be changed for an appropriate speaking exam. Students suggestions vary under many different topics. The most common suggestion is that the speaking exams should not be done at all or should be shorter. They suggest to have regular and shorter assessments which are done in the class during the term rather than having a speaking exam. Almost all of the students believe that the exam is beyond their knowledge and ability and should be simplified. They mostly state that the questions should be related to the things that they have been studying in the class. They think more activities are needed to be done before the exam and similar examples, topics or tasks should be studied in their listening and speaking classes.

Another common answer is that the format of the exam should be changed. The exam is suggested to include only interviews or presentations. They believe it would be a better way only to prepare presentations and talk about their presentations in the exam. Teachers should ask questions about their presentation in order to prevent memorization.

Another suggestion is about the interlocutors' behaviors and body language. In order to have a better atmosphere in the exam, the students suggest that the interlocutors should always smile, and be relaxed. By this way, they feel relaxed as well which will help their performances.

In addition to these one student suggests that more research and questionnaires should be done among students and they should be asked about what they want as an exam.

The last and the most interesting suggestion is from a student who thinks longer speaking exams are needed and that they are the only exams should be done. He suggests to remove all other kinds of exams and to have only speaking and writing exams. He believes, by this way, students will be more relaxed as they are not struggling with too many kinds of exams during the term. Furthermore, he adds that writing and speaking performances of the students are enough to assess their grammar, vocabulary knowledge and listening skills.

CHAPTER 5

5. CONCLUSION

In this chapter, a summary of the chapter is given firstly. Then, the results are displayed and discussed. Lastly, an assessment of the study and some implications for further research are given.

5.1. Summary of the Study

Zirve University Preparatory Program was chosen as the case for this study because speaking assessment procedure had regularly been changed during the last three years. Therefore, this study aimed to investigate whether the speaking component of the final exam at Zirve University Prep School has any problematic areas or not. The subjects of the research were ten teachers and fifteen students who had registered the program in September 2012. The researcher has also been an instructor at Zirve University since September, 2011.

This research was designed to look for the answer for how valid and reliable is the speaking exam held at Zirve University, specifically in level B. Level B was intentionally chosen as this group has the largest number of students. In order to find the answers to the research questions, interviews were conducted both with the students and the teachers; additionally, level B students‘ speaking exam scores were used. The interviews were analyzed in detail and common points from these interviews were used to gain data for interpretation of the speaking exam. Furthermore, the inter-rater reliability of the exam was investigated by using the speaking exam scores of level B students from both graders.

5.2. Discussion

This section discusses the outcomes of the study and draws conclusions about the research questions one by one.

The first research question is about the validity of the speaking component of the final exam at Zirve University Preparatory School. In order to find answer to this question a content validity interview was held with the Listening and Speaking Coordinator of the program.

It is known that in order to determine the content validity of an exam, the test's content should be examined to see if it "includes a representative sampling of what has been taught in a particular course, predetermined test objectives and test specifications" (Sak, p.89, 2008). As a result of this interview, it is found out that the speaking component of the final exam can not be regarded to have content validity. The exam can be named as an achievement test for the listening and speaking courses at Zirve University since it is done at the end of each term and used as a tool to determine whether the students are proficient enough to finish the level they have been studying and continue with the next one. Therefore, it is very important for such kind of a test to be related to the content of the course, that is to say, to be accepted to have content validity. Joughin (1998) clarifies this like “the validity of oral assessment is considered to be one of the strengths of this format” (p. 376).

However, it is also concluded that the students are familiar with the content of the speaking exam although the exam can not be considered to reflect the course content. This is achieved with personal efforts of the skill coordinator and Listening and Speaking course teachers. The teachers prepare similar activities and sample exam tasks the week before the final exam and have the students study and practice for the exam during last week of the term.

It is also found out that the exam not only lack of a representative sample of listening and speaking course content, but also is structured in a relatively high level than the level of the book students have been using as a course book. This is one of the most problematic aspects of the final speaking exam. From this viewpoint, one can conclude that the students take the final speaking exam only with one week’s preparation, although the exam is supposed to test a whole term.

In addition to these, it can be assumed that the main problem causing the exam not being valid in terms of content is that the institution does not have a speaking and listening curriculum. Thus, this also results in not having certain test specifications, so the speaking exam can not be constructed according to such aims, descriptors or standards.

However, the tasks and the materials used in the exam can be regarded as authentic and meaningful enough since real pictures are used in these tasks. Furthermore, the exam contains some kinds of speech acts like greeting, describing, expressing ideas, discussing, agreeing and disagreeing. Yet it is also accepted that, sufficient research is done neither for determining the desired test content nor for evaluating this test content.

Considering all these, we reach the answer for the first Research Question and conclude that speaking component of final exam at Zirve University Preparatory School does not possess the quality of content validity.

In order to investigate the answer to the second Research Question, which is about the inter-rater reliability of the exam, Pearson Product Moment Correlation Coefficients were calculated. According to the results, it can be concluded that the exam scores seem to have statistically high inter-rater reliability (see table 1). The correlation coefficients obtained from the graders is 0.941, which proves the fact that the inter-rater reliability of the exam is as high as preferably should be. As Mcnamara & Lumney (1993) suggest, one solution to the rating problems is to train the both native speaker interlocutors and the graders. From this point of view, this result can actually be accepted as an unexpected result since no training was done for the graders before the exam.

It can be said to be an evidence that there accepted to be a mutual relationship between validity and reliability, and problems related to such a quality of an exam may affect the other quality as well. As in this case, it is possible for an exam to be reliable but invalid too. Therefore, in order to eliminate such kind of inconsistencies and problems, special procedures should be applied before and during the exam, specifically if the exam is of subjectivity such as speaking exams. As a result, primary concern for test developers and the institutions should be trying to pay required attention to these issues of reliability and validity and increase the quality of their exams.

The third Research Question is asked to find out the ideas of the instructors who took part in the final speaking exam. Interviews were held for gathering data for this research question.

The first thing to be concluded is that speaking exam in this format is accepted to be better than the previous ones tried at Zirve University Preparatory Program by the instructors. One of the most important features of an assessment tool is to provide students with a variety of contexts in which to practice oral communication skills (Helsel & Hogg, 2006). Instructors' general feelings about the speaking final exam is in parallelism with this idea that they conclude with this exam format it is tried to assess the students' oral communications skills rather than their knowledge about the book.

About the format of the exam pairing is accepted to be the best way to assess the students' speaking ability. As having the students one by one causes a more stressed atmosphere, and increases the students' nervousness, it is not desired to be used by the teachers. Although there are some ideas about using different ways for different levels of English the mostly accepted way is pairing. It can also be concluded that pairing, which helps students feel more relaxed, is believed by the teachers to be a more fair and practical way for a speaking exam.

The task types used in the final speaking exam is another issue of the research. According to a study done by Huei-Chun (2007), students perform better and get higher scores for the task of answering questions than for that of picture description. The reason for this is stated in another study by Halleck (1995) as there are more opportunities for the students to answer questions in English than to describe pictures and make presentations in English. On the contrary, the results obtained from the interviews held with the instructors show that interview-like exam tasks or presentations are not considered as suitable task types for assessing speaking ability. Their reasons for this is that such kind of tasks result in students' memorization which also affects the grading negatively. In addition to these, it can be concluded from the interviews that the same task types, like discussions in this case, should not be used in all levels since students in low levels inevitably have difficulties in critical thinking skills.

About using visual aids in the assessment of speaking performance, it is concluded from the interviews that all of the instructors believe it has a positive and supporting effect on students' performances. Although the fact that choosing appropriate and variable pictures is of high importance, using them in a speaking exam give students the chance to have something to start and continue with.

It is understood from the answers of the instructors that they generally have no problems about the seating format of the exam. Some of the instructors mention that they need to see the students' faces and gestures in order to grade more fairly, while the others think it is needed not to be seen by the students during grading as this is believed to cause more stress for students.

In concern with being native or non-native in language teaching, Han (2004) states that the goal of L2 teaching should not be about creating native speakers, but rather L2 users. In line with this, it can be concluded from the interviews that most of the instructors believe being native or non-native has no relationship with grading or carrying out a speaking exam.

In a study, Samimy & Brutt-Griffler (1999, cited in Lasagabaster & Sierra, 2010) point out the fact that in many countries native speakers of English often lack adequate TEFL qualifications and gain entry into the profession or have positions at language academies simply because they are native speakers. Actually this is the case for Zirve University as well. In connection with this, some of the instructors also think that being non-native can not be regarded as a disadvantage but an advantage for a better speaking exam.

In another study, Davies (2011) supports this idea by saying “it’s not surprising when a group of educated non-native speakers outperforms a group of native speakers” (p. 306). A similar idea is also concluded from the interviews as well. The native instructors think that the only problem about non-native speakers is that they tend to focus on grammar and vocabulary while grading a speaking performance, and that with a sufficient amount of training it would be better.

Having different task types, especially discussions, in the exam; assessing the ability to speak in a real sense and preventing memorization are the aspects of the final speaking exam that are mentioned as mostly liked aspects. On the other hand, the duration of the exam, which is said to be too long and tiring, and the rubrics used by the graders are accepted as the negative and mostly disliked aspects of the final speaking exam.

In order to answer the Research Question four, which is about students’ ideas about the exam, interviews were held with fifteen Level B students. The questions in the interview are nearly the same with the one conducted with the instructors.

The first thing to be concluded from the answers of the students who are chosen as the subjects of the study is that nearly all of them believe that there is no need for having a speaking exam as they are always very nervous in the exam, so they are never able to show their speaking ability in a exam. The most important conclusion to be drawn from the answers, which also supports the finding for the Research Question one, is that they claim the topics and the tasks in the exam have no relationship with the ones they have been studying during the term in their listening and speaking classes. The tasks and the topics in the exam are said to be more difficult and unfamiliar for them. This indicates and supports the fact that the speaking final exam does not possess the quality of content validity.

In terms of the format of the exam, it can be concluded that on the one hand, half of the students agrees having the exam individually is better, and on the other hand, the other half believes having the exam with pairs is rather easy and comfortable. The ones who say being individual in the exam is better, state their reasons by saying that they do not want to be responsible for their friends’ performances, they are affected by their pairs negatively and it is more stressful not to be alone in the exam. On the contrary, the other half defends that being with their friends is something that helps them relax.

In relation with the exam tasks, it can be concluded that the students all agree on the issue of having presentations or interviews in the speaking exam is the best way for them. In his study, Huei-Chun (2007) concludes that answering questions is considered to be a kind of semi-interview, that's why this is perceived to be more stressful than other formats, and also states that picture descriptions are preferred by the students. However, the findings of the interviews do not support Huei-Chun's study, since almost all students think that having interviews and presentations are better rather than any other test formats.

Besides, the findings about using visual aids in the speaking exams indicate the same conclusion with Huei-Chun's study. He suggests (2007) that "it may be implied that the test takers seem to be more interested in the oral task of picture description because of the visual cues provided by the task " (p. 9). Similarly, the subjects of this research state that they like the idea of using pictures as it provides them with some starting points. However, one more thing to be concluded from the answers is that they believe it would be better if they have the chance to choose the picture to speak about; and also if their speech is not limited with the questions asked with the pictures.

About the grading procedure of the exam students seem to be content with the current situation. Apart from the fact that some of them believe the percentage of the score that the interlocutors give should be higher. One more conclusion to be drawn from the answers is that the grading procedure and having two other instructors during the exam to grade their performances make the students feel uncomfortable and more nervous. It is claimed to be a better way, if the graders do the grading for each performance later, by watching the videos. This claim indicates parallelism with one of the findings of a study done by Mulac & Sherman (1975). They suggest that using videotapes and grading from these have many advantages. One is "the avoidance of any rater influences upon the speech performances, since raters were not in the audience when the speeches were originally delivered" (p.141).

Lasagabaster & Sierra state in their studies that there are a number studies aiming to debate about comparing native speaker teachers and non-native speaker teachers; however, they mostly focus on teachers' opinions rather than the students' opinions. They give the conclusion that most generally the students prefer to have native teachers in an assessment situation (2002). A combination of native and non-native teachers is another preferred way, but not having non-native teachers alone. In connection with this, the results of the interview conducted with the students indicate that half of the students prefer to have native teachers in their exams, specifically their own speaking and listening teachers. Yet it can also be said that nearly the other half of the students state that they want Turkish teachers in the speaking exams since they believe they will understand them better. Although it is stated that they would also prefer Turkish teachers, they also accept that native speakers' scores are generally higher than theirs.

Another conclusion to be drawn from the interview is that interlocutor's body language is of vital importance for the students' performances. They all mention that interlocutor's body language, facial expressions and behaviors are all important factors that affect their performances during the exam.

The first task of the exam is said to be the most liked aspect of the final speaking exam; whereas, there are many issues about the exam that are disliked such as discussions, topics, duration, teachers' behaviors, being graded by three teachers and pairing.

5.3. Suggestions for a Better Speaking Exam Procedure

In order to find out the answer to the last Research Question, which is for seeking suggestions for a more appropriate exam procedure at Zirve University, the answers to the twelfth questions of both interviews are used. The results for suggesting a better speaking exam procedure can be listed as below:

- There should be more clarification about the shared responsibilities during the speaking exam. A leader is needed to be named specifically, in order to have a more organized exam procedure.

- The rubrics are needed to be revised and restructured in accordance with the speaking exam's practicality and reliability. Clearer instructions and descriptors are also needed.

- Exam tasks are also needed to be revised and changed. It is suggested that different task types may be used in different levels. Low levels should not have the same format with higher levels. Discussion task, specifically, should be reconsidered and changed. Furthermore, the questions which make the discussion more controlled, might be removed or changed in order to enable students to speak and comment on the topics freely.

- Rater and interlocutor training is accepted to be very crucial for a speaking exam to be fair, objective and reliable. Thus, more training sessions before the exam should be organized and teachers should be trained adequately and regularly.

- A curriculum, in real terms, particularly for listening and speaking, should be prepared within the shortest time. The institution should start to assign staff for this issue and work for a curriculum immediately. By this means, the inconsistencies between the speaking exam and the course content is believed to be eliminated.

- Fluency should be considered more during the speaking exam rather than accuracy. The instructors who are assigned to be graders should be explained and convinced about this issue.

- Listening and Speaking course books may be changed and the ones which are appropriate to critical thinking skills should be used.

- Pairing is accepted to be a good way, although it is suggested that the pairs should not be chosen randomly. Especially Listening and Speaking course teachers should work on the correct matching between the students more carefully.

- The students should be provided with more pictures during the exam and permitted to chose the pictures they want to speak about.

- Interlocutors' paraphrasing of the questions or leading the students might not be prohibited. Not always, but in some cases they may be permitted to help the students who need support.

- Instead of having speaking exams only at the end of each term, a kind of process assessment for Speaking and Listening course is another suggestion. Particularly speaking, since it is a productive skill, may be assessed during the whole term. Portfolios or in-class assignments can be considered as other possible ways to assess this skill.

- As discussion type tasks are regarded too difficult by almost all of the students, the task itself, the format and the procedure of using these tasks should be reconsidered. Perhaps some other kinds of tasks can be used instead of discussions.

- The instructors, particularly the interlocutors should be reminded about the issue that their body language is one of the most affective factors for students' performances and they need to be explained clearly how to behave during the speaking exams.

- More research on this subject should be done and the students should be asked about what they want and do not want more often.

5.4. Assessment of the Study

This research study aimed to find out the strengths and weaknesses of the final speaking exam at Zirve University Preparatory School. In spite of the fact that the findings and the results belong to this institution’s case and can not be generalized, the instruments and the procedures of the study can be used in other similar studies.

Using interviews as the instrument of the study can be stated as one of the limitations for the study, in terms of the time required. As having fifteen interviews with students and eleven interviews with the instructors take a great amount of time both for conducting and for analyzing them, it was difficult and time consuming for the researcher.

On the other hand, this study can be improved by using questionnaires. By this way, the number of the participants of the study can be increased; besides, it would be faster to gather data from these questionnaires.

Another way to improve the current study may be having interviews with the administration as well. Non of the participants was a member of the administration; having their ideas may have brought further insights to the results and the findings of the study.

Moreover, the content validity of the exam can have been sought not only by having an interview with the Speaking and Listening Coordinator, but also by asking the opinions of some instructors of Listening and Speaking course or some other staff from the administration.

5.5. Implications for Further Research

Joughin (1998) says that “apart from a small number of studies on anxiety, the literature does not include studies of oral assessment from the perspective of students“ (p. 376). Instead, many studies are based exclusively on teachers’ or educational researchers’ perspectives of oral assessment. Therefore, further research can be done in a more detailed way, in order to investigate the issue of speaking assessment from the students’ point of view. Studies of students’ descriptions and perceptions of speaking assessment most probably lead the subject to quite a different way.

This research study focused on only some types of the exam qualities like content validity and inter-rater reliability. Thus, another implication can be the fact that further research can be done in order to investigate other types of validity and reliability as well.

In addition, this study may be helpful for the researchers, teachers, institutions, test developers and other people who are interested in speaking assessment as it gives a model for further study and investigation of related area. Further studies can be conducted on a larger scale by any institution in order to investigate and evaluate their own circumstances for the assessment.

6. REFERENCES

Alderson, J. C., Clapham C., Wall D. (2005). Langıage Test Construction and Evaluation, United Kingdom, UK: Cambridge University Press.

Bachman, L. F. & Palmer, A. S. (1996). Language Testing in Practice. Oxford: Oxford University Press. Brown, H. D. (2001) Teaching by Principles. England, Pearson Longman Education.

Brown, J. D. (2005). Testing in Language Programs: New York, NY: McGrow-Hill.

Bygate, M. (2001). The Cambridge Guide to Teaching English to Speakers of Other Languages. Cambridge: Cambridge University Press.

Burges, S., Head, K. (2005). How to Teach for Exams. United Kingdom, UK: Pearson Education Press.

Caban H. L. (2003). Rater Group Bias in the Speaking Assessment of Four L1 Japanese ESL Students. *Second Language Studies, 21*, 1-44.

Carlson R. E., Howell D. (1995). Classroom Public Speaking Assessment: Reliability and Validity of Selected Evaluation Instruments. *Communication Education, 44*, 87-94.

Cohen, A. D. (1980). Assessing Language Ability in the Classroom. Boston: Heinle & Heinle Publishers.

Davies, A. (2011). Does Language Testing Need the Native Speaker? *Language Assessment Quarterly, 8* (3), 291-308.

Fulcher, G. (2003). Testing Second Language Speaking. London: Pearson Longman Education.

Güney, İ. (2010). An Investigation into the Causes of Speaking Problems Experienced by Learners of English at Tertiary Level, unpublished MA thesis, Çanakkale Onsekiz Mart University, School of Social Sciences, Çanakkale.

Hager P., Gonczi A., Athanasou J. (1994). General Issues about Assessment of Competence. *Assessment and Evaluation in Higher Education, 19* (1), 3-15.

Halleck, G. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. *Modern Language Journal*, *79*, 223-34.

Han, Z. H. (2004). To be a Native Speaker Means not to be a Nonnative Speaker. *Second Language Research*, *20*, 166–187.

Heaten, J. B. (1990). Writing English Language Tests. London: Longman.

Helsel, C. R., Hogg M. C. (2006). Assessing Communication Proficiency in Higher Education: Speaking Labs Offer Possibilities. *International Journal of Listening, 20* (1), 29-54.

Henning, G. (1987). Guide to Language Testing. Cambridge: NewBury House Publishers.

Huei-Chun, T. (2007). A Study of Task Types for L2 Speaking Assessment. *Annual Meeting of the International Society for Language Studies.* Honolulu: HI.

Hughes, A. (2003) Testing for Language Teachers. United Kingdom, UK: Cambridge University Press.

Joughin G., (1998). Dimensions of Oral Assessment. *Assessment and Evaluation in Higher Education, 23* (4), 367-376.

Kim H. J. (2006). Issues of Rating Scales in Speaking Performance Assessment. *Teachers College, Colombia University Working papers in TESOL & Applied Linguistics, 6* (2), 1-3.

Kitao S. K., Kitao K. (1996). Testing Speaking. Retrieved from: http://www.eric.ed.gov/PDFS/ED398261.pdf : ED 398 261.

Lasagabaster D., Sierra J. M. (2002). University Students' Perceptions of Native and Non-native Speaker Teachers of English. *Language Awareness, 11* (2), 132-142.

Luoma, S. (2004). *Assessing Speaking.* United Kingdom, UK: Cambridge University Press.

McNamara, T. F. (1996). Measuring Second Language Performance. London: Longman.

McNamara T. F., Lumney T. (1993). The Effect of Interlocutor and Assessment Mode Variables in Offshore Assessments of Speaking Skills in Occupational Settings: *Language Testing Research Colloquium.* England, University of Cambridge.

Mead, N. A., (1980). Assessing Speaking Skills: Issues of Feasibility, Reliability, Validity and Bias. *Annual Meeting of the Speech Communication Association,* New York: NY.

Nakamura, Y. (1996). Assessment of English Speaking Ability. *Journal of Humanities and Natural Sciences*, 102, 25-53.

Mulac A., Sherman A. R. (1975). Relationships among Four Parameters of Speaker Evaluation: Speech Skill, Source Credibility, Subjective Speech Anxiety, and Behavioral Speech Anxiety. *Speech Monographs, 42*, 302-310.

Nunan, D. (2002). Research Methods in Language Learning. Cambridge: Cambridge University Press.

O'Sullivan B. (2008). Notes on Assessing Speaking. retrieved from: http://lrc.cornell.edu/events/past/2008-2009/papers08/osull1.pdf

Önal, A. (2010). Testing and Assessment of Speaking Skills in Preparatory Classes, unpublished MA thesis, Selcuk University, School of Social Sciences, Konya.

Rubin R. B., (1982). Assessing Speaking and Listening Competence at the College Level: The Communication Competency Assessment Instrument. *Communication Education*, *31* (2), 19-31.

Sak, G. (2008). An Investigation of the Validity and Reliability of the Speaking Exam at a Turkish University. unpublished MA thesis. Middle East Technical University, School of Social Sciences, Ankara.

Salaberry, R. (2000). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing, 17* (3), 289-310.

Stiggins, R. J., Backlund, P. M., & Bridgeford, N. J. (1985). Avoiding Bias in the Assessment of Communication Skills. *Communication Education, 34*, 135-145.

Taylor, L. (2003). The Cambridge Approach to Speaking Assessment. Retrieved from: http://www.CambridgeEsol.org/researchnotes

Underhill, N. (1987). Testing Spoken Language. Cambridge: Cambridge University Press.

Ur, P. (1996). A Course in Language Teaching. Cambridge: Cambridge University Press.

Weir, C. J. (2005). Language Testing and Validation. New York, NY: Palgrave Macmillan.

Weir, C. J. (1990). Communicative Language Testing. New York, NY: Prentice Hall.

7.APPENDICES

7.1. APPENDIX A: Guidelines for Final Speaking Exam for Level B

Guidelines for Interlocutors

Task 1: (student/interlocutor tasks)

Interlocutors should...

- make eye contact and try to make the students comfortable;

- encourage the students to speak as much as possible during their introduction. As an example, if the students mention less than three of the topics on the prompt sheet, the interlocutor could say "Anything else? Your family?";

- ask different questions to each student;

-repeat a question if the student does not understand;

-select another question from the list if the student still does not understand;

-NOT paraphrase the given questions.

Tasks 2&3: (student/student tasks)

Interlocutors should intervene if...

-the students do not understand what is required of them;

-there is a communication breakdown;

-there is a major imbalance between the two students' contributions;

-the amount of language produced by the students is insufficient;

-the students are speaking in any language other than the target language.

In such cases the interlocutor should...

-repeat all or part of the task instructions;

-paraphrase all or part of the task instructions;

-invite the students to talk about one specific aspect of the task;

-invite the candidate whose contributions seem to be unsatisfactory (i.e. too short or incomplete) to talk about one specific aspect of the task or elaborate on something s/he said;

-politely ask that the students use only the target language.

Guidelines for Examiners

Examiners should...

-sit behind the students (this is to reduce student stress)

-NOT speak to, or translate for, the students

-NOT make eye contact with the students;

-NOT stand up, walk around, work on their computer/cell phone, or draw any attention to themselves in any way;

-be familiar with the follow up questions and speaking prompt;

-be familiar with the CEF (Common European Framework) in order to better understand and assess each students true level and ability (Level A=A1, Level B=A2, Level C=B1, Level D=B2);

-be familiar with the written rubric;

-grade each students based on his/her own performance, without comparing against the other students performance

-grade each student individually, WITHOUT conferring with the other examiner or interlocutor

7.2. APPENDIX B: Final Speaking Exam Tasks

Sheet 1 / Task 1 Speaking Prompt

Name?

Age?

Hometown?

Job?

Hobbies?

Family?

Sheet 2 / Task 2 Prompt Sheet 1

CEMYILMAZ
COMEDY SCHOOL

CEMYILMAZ

22, Ataturk Street

We'll teach you to act, perform stand up and tell jokes

Classes 12 to 5 p.m. Every Sunday

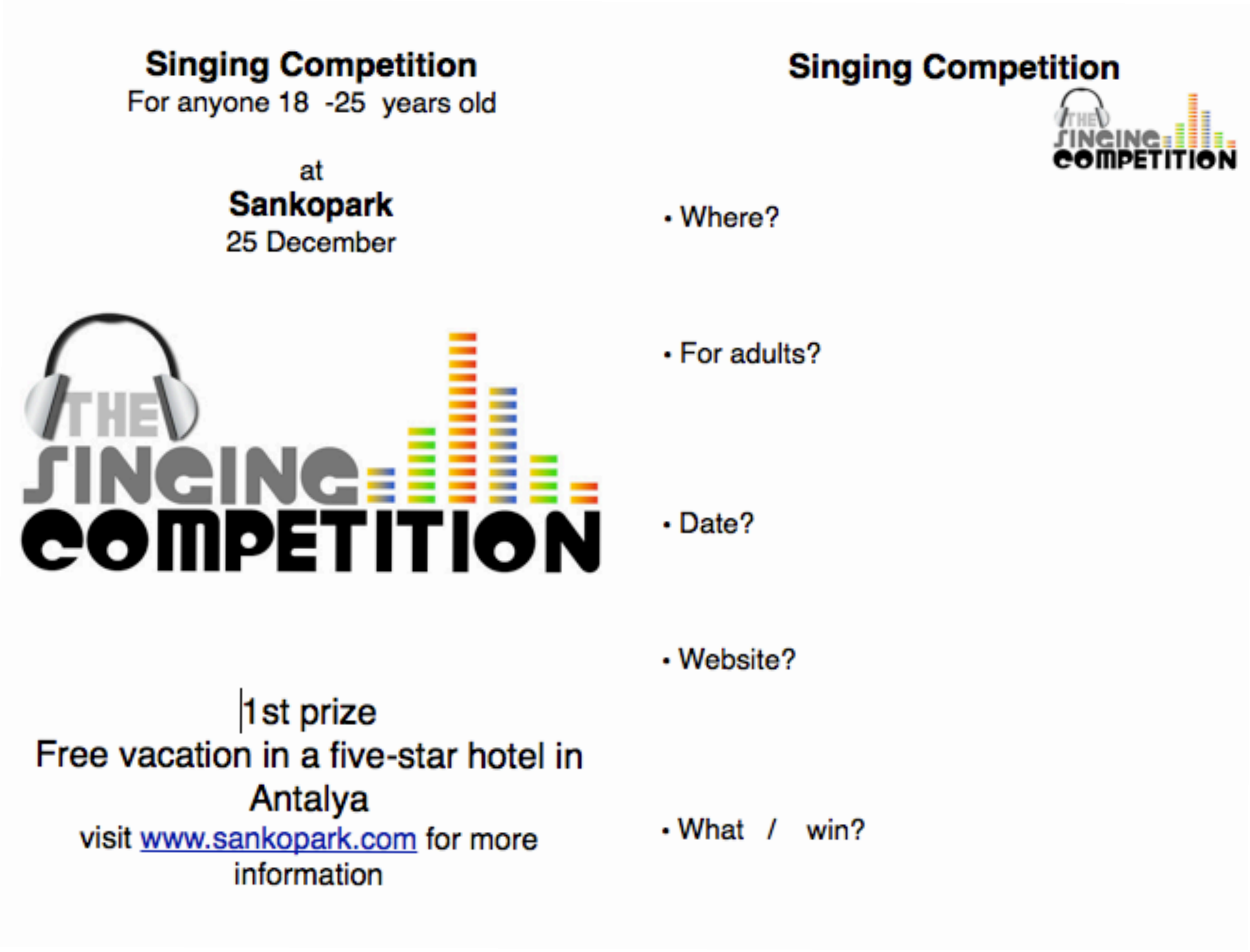
Fee: 200 TL a month

Visit: www.cmylmz.com

Comedy School

- Name / School?
- What / learn ?
- When / classes?
- Cost ?
- address ?

Sheet 3 / Task 2 Prompt Sheet 2



THE
SINGING
COMPETITION

Singing Competition
For anyone 18 -25 years old

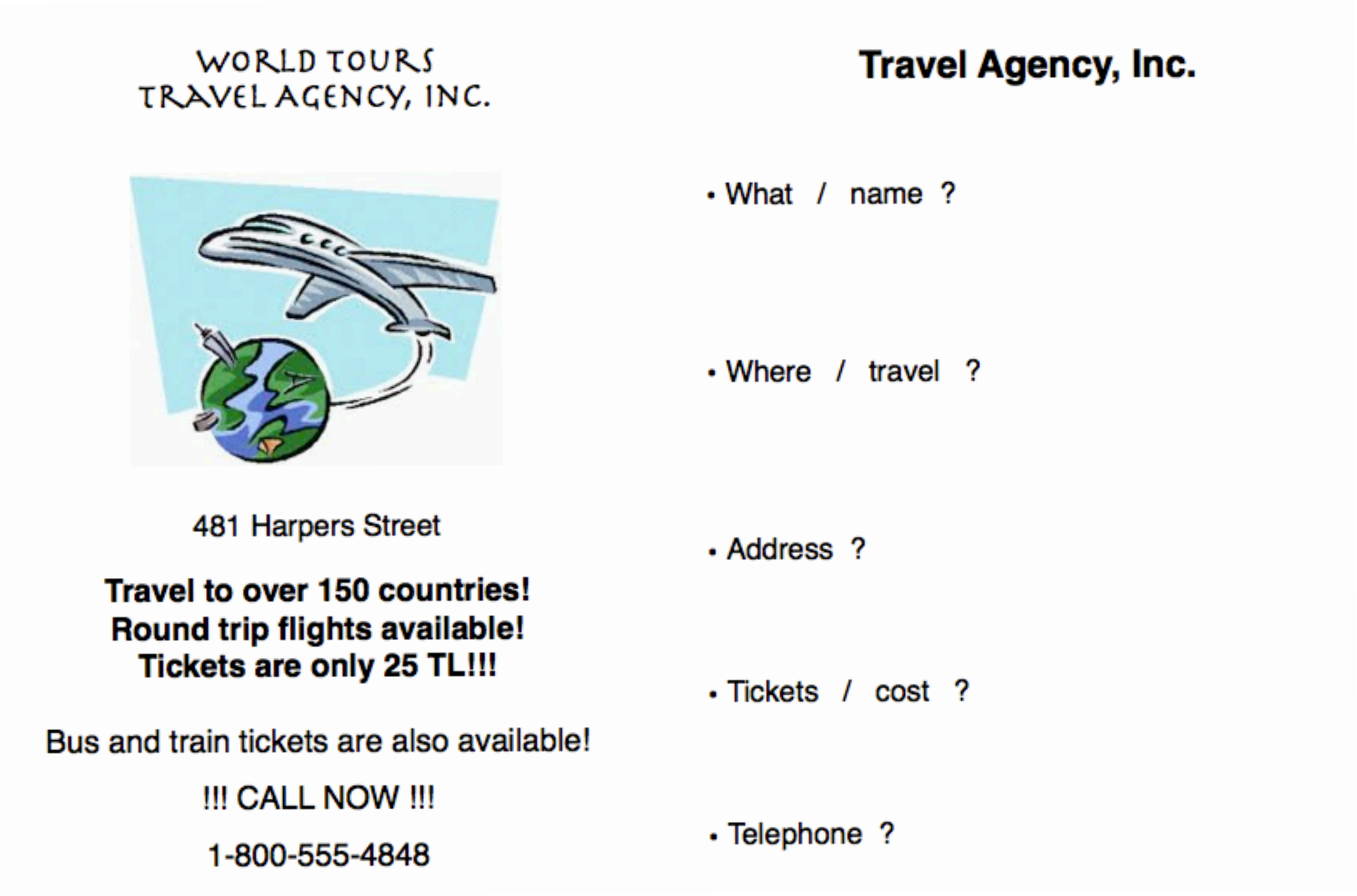
at
Sankopark
25 December

1st prize
Free vacation in a five-star hotel in Antalya
visit www.sankopark.com for more information

Singing Competition

- Where?
- For adults?
- Date?
- Website?
- What / win?

Sheet 4 / Task 2 Prompt Sheet 3



WORLD TOURS
TRAVEL AGENCY, INC.

481 Harpers Street

Travel to over 150 countries!
Round trip flights available!
Tickets are only 25 TL!!!

Bus and train tickets are also available!

!!! CALL NOW !!!

1-800-555-4848

Travel Agency, Inc.

- What / name ?
- Where / travel ?
- Address ?
- Tickets / cost ?
- Telephone ?

Sheet 5 / Task 2 Prompt Sheet 4

NEW YORK FIGHTERS

Basketball Game

New York Fighters vs. Chicago Extreme

@ the New York Sports Arena

Tickets On-Sale: Students = 38TL
Normal Ticket = 54TL

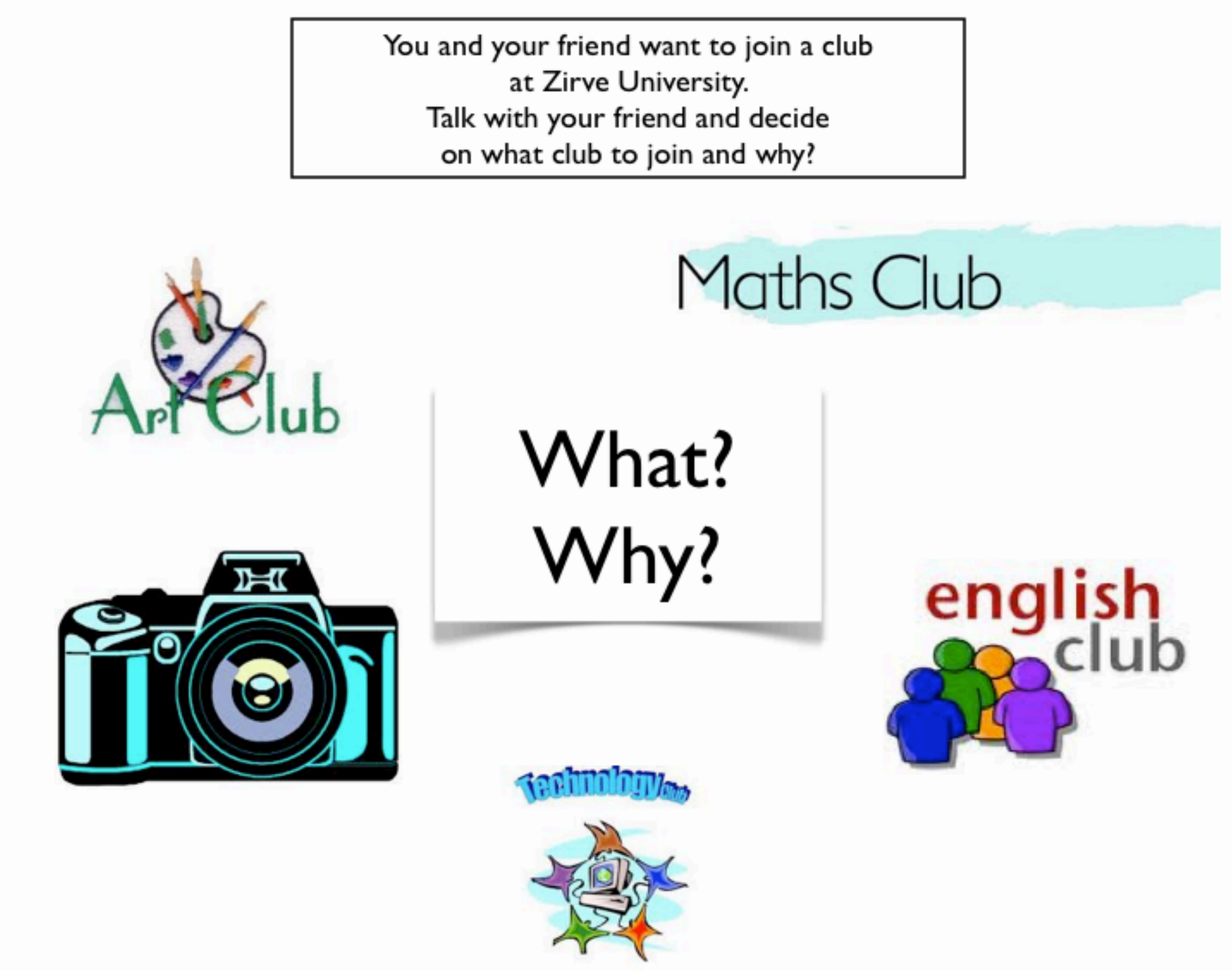
For more information call:
1-888-212-8888

Buy online tickets at www.buytickets.com

Basketball Game

- What / teams ?
- Where / playing ?
- Student ticket / cost ?
- Telephone ?
- Tickets / online ?

Sheet 6 / Task 3 Discussion Task 1



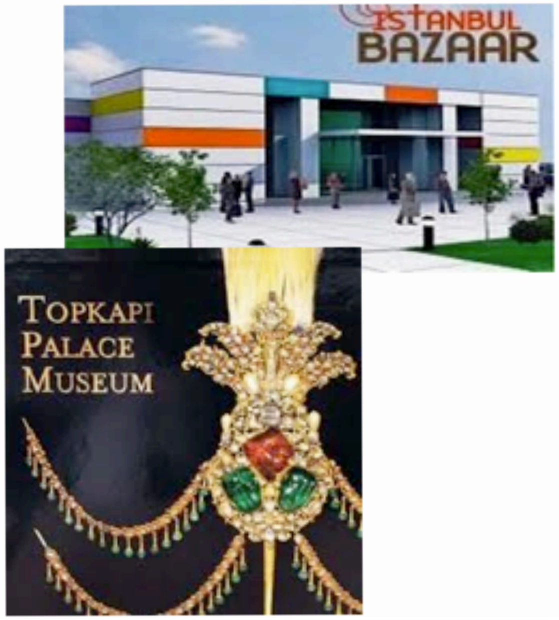
You and your friend want to join a club
at Zirve University.
Talk with your friend and decide
on what club to join and why?

What?
Why?

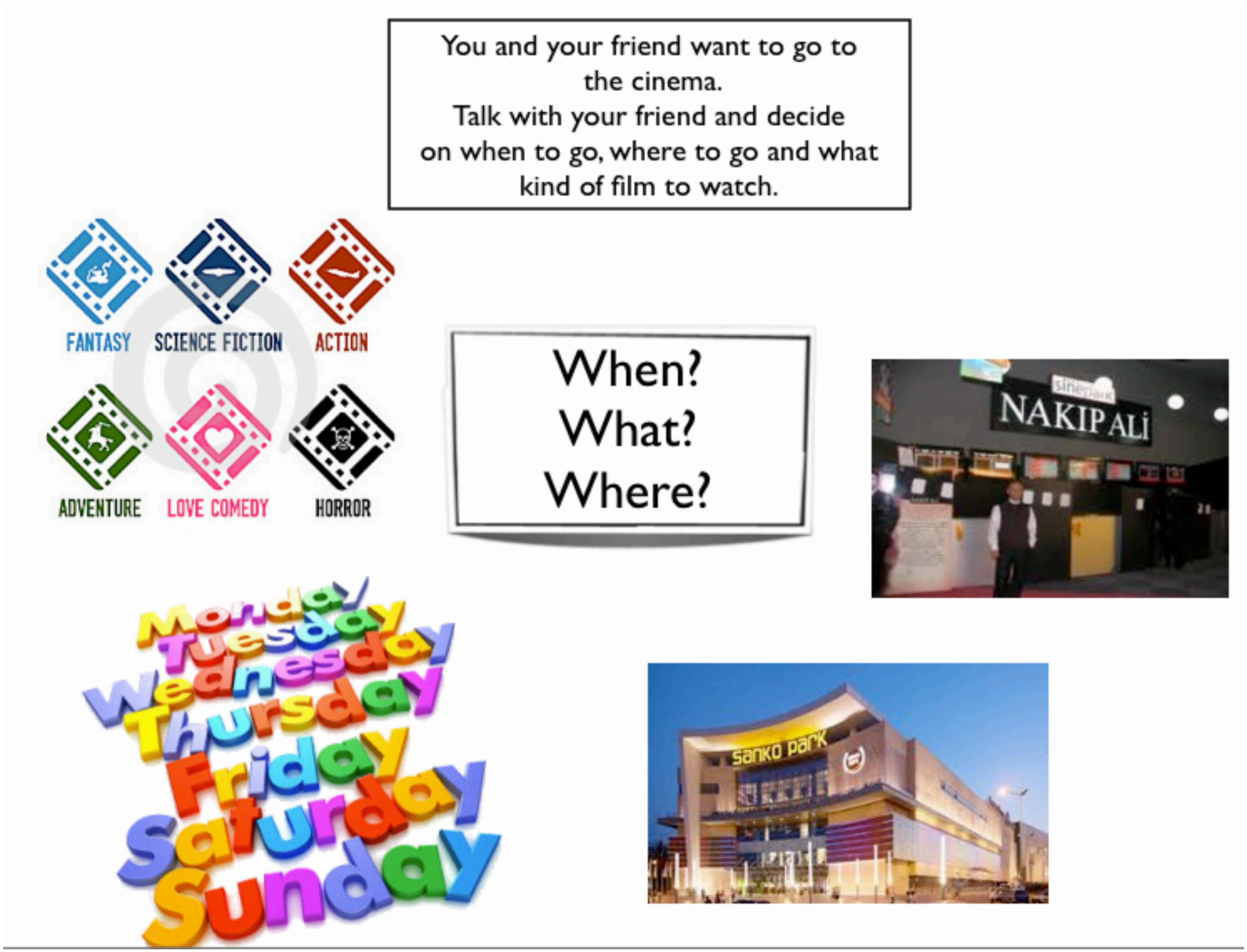
Sheet 7 / Task 3 Discussion Task 2

You and your friend want to go to İstanbul. Talk with your friend and decide on how to go , what to do and what you will eat there.

What?
How?
When?



Sheet 8 / Task 3 Discussion Task 3



You and your friend want to go to the cinema.
Talk with your friend and decide on when to go, where to go and what kind of film to watch.

FANTASY SCIENCE FICTION ACTION

ADVENTURE LOVE COMEDY HORROR

When?
What?
Where?

Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

Sheet 9 / Task 3 Discussion Task 4

You and your friend want to buy a present for
your teacher on the Teacher's Day.
Talk with your friend and decide
on what present to buy.

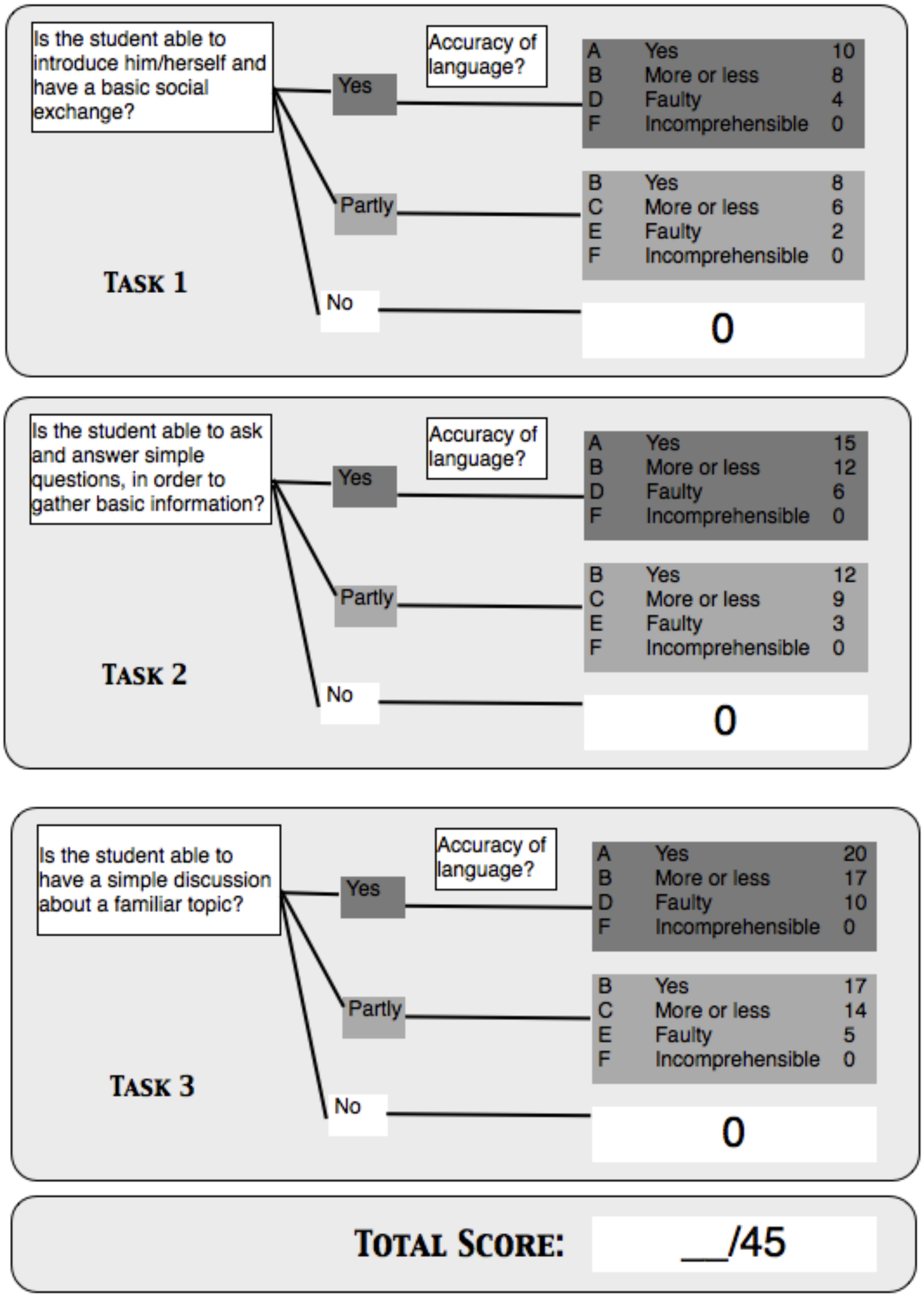
What?
Why?

????

7.3. APPENDIX C: Level B Interlocutor Holistic Scale

B	Global Achievement Scale
5	Handles communication of very common topics, despite hesitation. Can maintain conversation, although inaccuracies may occur.
4	*Performance shares features of Bands 3 and 5*
3	Handles only basic conversation in only very common situations. Most utterances tend to be very short - words or phrases - with frequent hesitation and pauses.
2	*Performance shares features of Bands 3 and 1*
1	Has difficulty conversing even about very familiar everyday topics. Responses are limited to short phrases or isolated words with frequent hesitation and pauses.
0	*Performance below Band 1*
	Total Score: __/5

7.4. APPENDIX D: Level B Graders' Rubric



TASK 1

Is the student able to introduce him/herself and have a basic social exchange?

Accuracy of language?

Yes

A	Yes	10
B	More or less	8
D	Faulty	4
F	Incomprehensible	0

Partly

B	Yes	8
C	More or less	6
E	Faulty	2
F	Incomprehensible	0

No

0

TASK 2

Is the student able to ask and answer simple questions, in order to gather basic information?

Accuracy of language?

Yes

A	Yes	15
B	More or less	12
D	Faulty	6
F	Incomprehensible	0

Partly

B	Yes	12
C	More or less	9
E	Faulty	3
F	Incomprehensible	0

No

0

TASK 3

Is the student able to have a simple discussion about a familiar topic?

Accuracy of language?

Yes

A	Yes	20
B	More or less	17
D	Faulty	10
F	Incomprehensible	0

Partly

B	Yes	17
C	More or less	14
E	Faulty	5
F	Incomprehensible	0

No

0

TOTAL SCORE: __/45

TASK 1

Is the student able to introduce him/herself and have a basic social exchange?

Yes	Maintains a simple dialogue and exchange. Requires very little prompting and support.
Partly	With some difficulty, maintains a simple dialogue and exchange. May require some prompting and support.
No	Cannot maintain a simple dialogue and exchange, even with prompting and support.

Is the quality of the language adequate?

Yes	Is mostly intelligible, and has some control of phonological features at both utterance and word levels. Shows a good degree of control of simple grammatical forms. Uses a range of appropriate vocabulary when talking about familiar topics.
More or less	Is mostly intelligible, despite limited control of phonological features. Shows some control of simple grammatical forms. Uses appropriate vocabulary to talk about familiar topics.
Faulty	Has limited control of phonological features and is only sometimes unintelligible. Shows only limited control of a few simple grammatical forms. Uses a vocabulary of isolated words and phrases.
Incomprehensible	The student's performance contains so many errors that communication is (almost) impossible.

TASK 2

Is the student able to ask and answer simple questions, in order to gather basic info?

Yes	Can formulate simple questions with occasional grammatical errors. Produces responses which are extended beyond short phrases, despite hesitation.
Partly	Has some difficulty formulating grammatically simple questions. Responses are generally short phrases. May require some prompting and support. Has difficulty formulating questions. Responses are strained and often one-worded. Requires prompting and support.
No	Cannot formulate simple questions. Responses are often strained and one-worded, despite prompting and support.

Is the quality of the language adequate?

Yes	Is mostly intelligible, and has some control of phonological features at both utterance and word levels. Shows a good degree of control of simple grammatical forms. Uses a range of appropriate vocabulary when talking about familiar topics.
More or less	Is mostly intelligible, despite limited control of phonological features. Shows some control of simple grammatical forms. Uses appropriate vocabulary to talk about familiar topics.
Faulty	Has limited control of phonological features and is only sometimes unintelligible. Shows only limited control of a few simple grammatical forms. Uses a vocabulary of isolated words and phrases.
Incomprehensible	The student's performance contains so many errors that communication is (almost) impossible.

TASK 3

Is the student able to have a simple discussion about a familiar topic?

Yes	Maintains a simple exchange. Requires very little prompting and support.
Partly	With some difficulty, maintains a simple exchange. May require some prompting and support.
No	Cannot maintain a simple exchange, even with prompting and support.

Is the quality of the language adequate?

Yes	Is mostly intelligible, despite limited control of phonological features. Shows some control of simple grammatical forms. Uses appropriate vocabulary to talk about familiar topics.
More or less	Has limited control of phonological features and is only sometimes unintelligible. Shows only limited control of a few simple grammatical forms. Uses a vocabulary of isolated words and phrases.
Faulty	Has a very limited control of phonological features and is mostly unintelligible. Shows no control of simple grammatical forms. Has a very limited vocabulary of basic words.
Incomprehensible	The student's performance contains so many errors that communication is (almost) impossible.

7.5. APPENDIX E: Content Validity Interview Questions

1. Who is the test designed for? What is it designed for?
2. What is the basis for considering whether the test is appropriate to your students?
3. Do you have any test specifications?
4. Is test content relevant to test specifications?
5. Do the items or tasks in the test match what the test as a whole is supposed to assess?
6. Does the test produce a good sample of the contents of the syllabus of the preparatory class?
7. How well do tasks/ items of the test reflect the characteristics of speaking ability?
8. What research was conducted to determine desired test content?
9. What research was conducted to evaluate test content?
10. Are the tasks and topical contents relevant to the target language use domain namely, the potential uses, or the situations that the test taker is likely to encounter)?

7.6. APPENDIX F: Interview Questions

INTERVIEW FORM

PART 1

1. How long have you been teaching Listening and Speaking classes?
2. How many hours do you teach a week?
3. How many classes and students do you teach?
4. Of the seven hours devoted to Listening and Speaking courses for one class a week, how much do you usually allocate to teaching speaking?

PART 2

1. What do you think about the procedure of the final speaking exam procedure in general?
2. What do you think about the “time” issue in the final speaking exam? Is the amount of time allocated for each student enough?
3. Which one do you think is better for a speaking exam; having the students one by one, in pairs or in groups? What are your reasons?
4. What do you think about the format of the speaking exam? Which one would be better; interviews, presentations, discussions, role plays ect? Do you have any suggestions for the format of the exam?
5. What do you think about the importance of using visuals aids in the speaking exam?
6. What do you think about the grading procedure of the speaking exam?
7. What do you think about the physical circumstances (exam layout, seating etc.) during the speaking exam?
8. Do you think it is better to assign only native speakers as interlocutors and Turkish staff as graders? What do you think about examining your own speaking-listening class?
9. (For instructors) Do you pay attention to include some aspects of delivery such as body language, eye contact etc. into the assessment criteria for speaking?

9. (For students) Are you affected by the body language of the interlocutor? How?

10. What is the aspect of the final speaking exam you liked most?
11. What is the aspect of the final speaking exam you disliked most?
12. How could the test be improved? Do you have any suggestions on how to improve the procedure of ‘assessment of speaking’?