American University of Armenia

Yerevan

ASSESSING SPEAKING IN ARMENIAN EFL CLASSROOMS

A thesis submitted in partial satisfaction
of the requirements for the degree Master of Arts
in Teaching English as a Foreign Language

by

Nvard Grkikyan

2005

The thesis of Nvard Grkikyan is approved

ENGLISH DEPT.

Marianne Celce-Murcia

Marianne Celce-Murcia

E. Kekejian

Elisa Kekejian

Jo Lewkowicz, committee chair

American University of Armenia

2005

Acknowledgments

I would like to express my gratitude to all the people who supported me while carrying out this research. My special thanks go to my supervisor Dr. Jo Lewkowicz for her useful comments, assistance and patience throughout the process of writing the thesis. I would like to thank also the Dean of the Department of English Programs Professor Marianne Celce-Murcia and the Assistant Dean Elisa Kekejian for their understanding and assistance in accomplishing the work. My special thanks go to the two teachers Gayane Petrosyan and Sona Iskandaryan who agreed to participate in the study and share their precious ideas. I am grateful also to their students. It would be impossible to realize the research study without their participation. My special thanks to my family and my friends for their moral support. It would be difficult for me if I did not have them sharing all the anxiety that I went through while writing the thesis.

Abstract

The purpose of the following study was to develop a rating scale for assessing students' oral performance in the fifth grade of Armenian secondary schools and test its usefulness and effectiveness in everyday classroom use. The research was carried out to investigate whether the assessment of speaking at secondary school level could be more effective and less subjective if guidelines and criteria were provided for teachers to assess their students. The data was collected through observations, along with recording of the lessons, which served as a basis for constructing the rating scale for assessing speaking. The next stage of the research was the development of the rating scale and its use in classrooms. The last stage of the study was the interview, which aimed at getting feedback from the two teachers who used the rating scale in their classrooms. The purpose of the observations and the recording of the lessons was to establish daily oral tasks and the aspects of language to be included in the rating scale. The effectiveness of the developed rating scale was tested by using it in classrooms for assessing students' oral performance. The purpose of the interview was to get feedback from the two teachers who used the scale in their classrooms. The steps taken during the research helped ascertain that the use of a rating scale with guidelines and explanations for each score and category is effective and advantageous in several ways: it fosters learning, includes students in the assessment process, addresses different aspects of spoken language and minimizes subjectivity in assessing students. The students were not interviewed; however, the teachers reported that they liked both the assessment process and the fact that they knew the basis upon which they were assessed.

Table of Contents

Acknowledgments iii

Abstract iv

Chapter One: Introduction 1

Chapter Two: Literature Review 4

2.1 Assessment 4

2.2 Nature of oral performance 12

2.3 Assessment of oral performance 13

2.4 Rating scales 14

2.5 Rating scales for classroom use 19

2.6 The case in Armenia 20

Chapter Three: Methodology 23

3.1 A Descriptive study 23

3.2 The process of data collection 23

Chapter Four: Results and Discussion 29

4.1 The observations 29

4.2 The interviews 32

4.2.1 Advantages of the rating scale 32

4.2.2 Difficulties in using the rating scale 34

4.2.3 Students' reaction to using the rating scale 35

4.2.4 Recommendations for changes to the rating scale 36

4.2.5 Summary of the interviews and findings 38

4.3 Distribution of scores 38

4.4 Implementation of the rating scale 42

Chapter Five: Conclusion **45**

5.1 Limitations of the study 45

5.2 Suggestions for further study 46

5.3 Contribution of the research 46

References **48**

Appendixes **52**

Appendix A **52**

Appendix B **54**

Appendix C **55**

Appendix D **58**

Appendix E **61**

Chapter One: Introduction

The system of assessment in Armenian secondary schools is not well-developed. Scoring of all tasks ranges from one to five, of which 'one' is rarely used by the teachers, generally only for punishment purposes. 'Two' is also rarely given: it denotes a fail and if a student's overall grade for the year is 'two' then s/he cannot pass on to the next year. 'Five' is also an exceptional grade because only excellent students are awarded this grade. Therefore, the range remains between 'three' and 'four'. In addition to the limited range of scores used, there is also another problem, one that concerns the explanation of those scores. No guidelines for any score exist. Teachers grade students based on their overall impression of the performance. Here the factor of fairness is obviously at stake since there is no guarantee that all test takers/students are treated equally (Bachman and Palmer, 1996). Subjectivity may be part of each assessment occasion. This is a particular problem for language classes where so many factors need to be addressed while assessing the students' ability to use the language (Carraquilo, 1994, cited in Shumin, 2002). The existing grading system therefore does not seem to be the best.

A baseline study conducted in 2002 aimed at investigating different aspects of teaching English in Armenian secondary schools. A whole chapter in Harutyunyan (2002) is devoted to assessment and examinations. There are two major problematic issues that the researcher raises concerning assessment. The first point that she notes is that most of the activities assessed in the classroom by the majority of teachers are activities testing the students' oral performance, which form the basis for the final grade for students' everyday performance, which in turn serves as a basis for the end of year grade. This is sub optimal because other skills are not considered. The student, for example, may master writing or reading skills first then listening and speaking. Therefore, basing the grade on the oral performance may in many cases mean that the teacher penalizes the student for failing to master this skill while at the same time not giving credit for another skill, for example reading, writing or listening. The other, and the most important part of the study, refers to the fact that there are no guidelines for teachers to base their assessment on (Harutyunyan, 2002), which is another main concern because again the factor of fairness is at stake. The degree of subjectivity in assessing students under these conditions may be very high. The second issue in its turn may lead to another problem when assessment across schools is considered. A student may get a 'five' in one school and a 'four' or a 'three' in another based on the standards that the schools and the teachers have. Consequently it appears that even though the same range of scores is used countrywide, each school may view and interpret it differently. For example, "a '5' in Armenian Language in an Armenian school might be different from a '5' in Armenian Language in a Russian school. The same is true for all other marks" (Muradyan, 1999, p. 22)

The baseline study gives the picture of the situation but no steps have as yet been taken for finding a way to solve the problem. There has been no follow up study which could suggest a way forward. This is of great concern because it may result in serious consequences for students' future education and also in their lives. Though classroom tests are considered to be low-stakes tests the results of which do not affect students' immediate futures (Davidson et al, 1997, cited in Clapham, 2000), they should not be treated indifferently because students form their understanding of what education is and start appreciating it from school. Therefore, a negative approach towards education developed in school may affect students' future lives (Clapham, 2000).

This study aims at filling the existing gap by suggesting a new rating scale for assessing students' oral performance for everyday classroom use. Only oral performance is addressed here as teachers base their final grades mainly on speaking activities (Harutyunyan, 2002). The new scale will include different aspects of oral communication with scores ranging from one to five for each category. The new rating scale will also be used in actual lessons for assessing students to test the usefulness and effectiveness of using such an assessment instrument. No changes will be made in the range of the scores in this rating scale but each category relevant to oral performance will be addressed separately which intends not to penalize students for what they do not know but give credit for what they know.

Therefore the research question is:

How effectively does a rating scale with criteria and explanation for each score reflect students' ability to speak in the target language in the fifth grade of Armenian secondary schools?

This question will be addressed throughout the research, the coming sections of which are: literature review, methodology, results and discussion and conclusion.

Chapter Two: Literature Review

This chapter reviews prominent studies relating to the present research. The main areas to be addressed are (1) the nature of assessment and its components: reliability, validity, authenticity, interactiveness, impact and practicality; (2) the nature of oral performance; (3) assessment of oral performance and (4) the use of rating scales for assessing oral performance. This will provide the background for discussing the way oral performance is assessed in Armenia and the possible consequences that may occur as a result of a lack of having a set way of assessing oral performance.

2.1 Assessment

Assessment, as defined by Clapham (2000), "is used both as an umbrella term to cover all methods of testing and assessment, and as a term to distinguish 'alternative assessment' from 'testing'" (p. 155). Tests, as Brindley (2001) points out, are administered on a 'one-off' basis. Assessment includes not only tests but also other methods of measuring students' learning and therefore is an on-going process. There are some other views concerning the differentiation between testing and assessment according to which 'testing' in the field of linguistics is mostly used to refer to large-scale standardized tests and 'assessment' is used to refer to more informal and school-based tests (Clapham, 2000). Though the author draws such distinction between these two terms as an accepted view by other researchers, she herself does not see any differentiation between them and uses the two terms interchangeably.

A distinction between assessment and evaluation is also made by different researchers. Assessment is considered a means of gathering information and making relevant inferences about language learners' knowledge and ability to use the language. However, evaluation mainly refers to the program and it deals with collecting information and making judgments about the quality of the whole program (Genesee and Upshur, 1996). Therefore, assessment deals with individual student learning while evaluation deals with other aspects of the teaching learning process as well. To summarize these three terms: "assessment subsumes testing and is, in turn, subsumed by evaluation" (Nunan, 2004, p. 138).

Many researchers believe that assessment has a great influence on students' learning process (Black and William 1998; Stiggins 2002, in Berry 2003); therefore, considerable attention should be paid to that aspect of language teaching. Moreover, Smith and Ragan (1999) point out that students' assessment should be based on the instruction and the learners' performance. Assessment should be used to make relevant changes to instruction. Bostwick and Gakuen (1995) share this opinion stating that assessment is a tool for making improvements in language teaching and also for having students be in charge of their own learning. They also emphasize the fact that assessment can only be used in this way if it is based on the curriculum of the program and if it is authentic (cited in Sook, 2003), that is the features of the test task correspond with features of the target language use (Bachman and Palmer, 1996).

In the process of test development the issue of test usefulness is essential for constructing tests which may give relevant information about students' real ability. Bachman and Palmer (1996) describe test usefulness "as a function of several different qualities, all of which contribute in unique but interrelated ways to the overall usefulness of a given test" (p.18). Therefore, while constructing language tests, considerable attention should be paid to those aspects of the test. Test usefulness, according to Bachman and Palmer (1996) is determined by construct validity, reliability, authenticity, interactiveness, impact and practicality. Each will be considered in turn in the following sections.

2.1.1 Validity

A test is said to have the characteristic of validity if "it measures accurately what it is intended to measure" (Hughes, 1989, p. 22). Construct validity, an essential quality for test usefulness, and its components, content and face validity, will be addressed here as important aspects of an achievement test.

Construct validity, which Bachman and Palmer (1996) consider one of the important qualities of test usefulness, deals with the interpretation made based on the scores, it is "used to refer to the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure" (p.21). That is, construct validity concerns the extent to which the score obtained on a test can be generalized to make valid judgments about the test taker's knowledge of the language.

A test of spoken language in the classroom is usually an achievement test, so it has to have also content validity and face validity which are considered as levels of construct validity (Bachman and Palmer, 1996). As content validity refers to the extent the course content matches with the test content (Bachman, 1990), secondary school language teachers have to develop tests based on the content which the students have covered during the course. Face validity of a classroom spoken test, which refers to whether the test "looks as if it measures what it is supposed to measure" (Hughes, 1989, p. 27), is also important as a test with face validity may motivate students to speak. The reason is that the test does not leave the impression that it is very difficult, which would make students think that it is beyond their ability to complete the task.

2.1.2 Reliability

While constructing language tests particular attention should be paid to how the tests are administered and scored so that "scores actually obtained on a test on a particular occasion are likely to be very similar to those which would have been obtained if it had been administered to the same students with the same ability, but at a different time. The more similar the scores would have been, the more reliable the test is said to be" (Weir, 1993, p. 29). A very important concern for having test reliability is to have inter-rater reliability and intra-rater reliability. In the first case the researchers need to correlate the scores assigned by different raters to the same test. Raters who rate large-scale tests have to have agreement among the scores which have been assigned following certain guidelines. In the latter case, the concern is about the consistency of the same rater in rating the same work on different occasions (Bachman, 1990). However, in classrooms inter-rater and intra-rater reliability are rarely of concern, as the teacher is the only one who is responsible for the scores assigned to the students; and the rating of any work is done only once.

Though both reliability and validity are important characteristics of a good test, sometimes validity is given more consideration than reliability. For example, in certain cases, if a choice has to be made, validity is given more importance for assessing speaking (Bachman, 1990). This is because of the complicated nature of setting test reliability. Reliability deals with identifying the source of the error and its effect on the score and "in order to identify sources of error, we need to distinguish the effects of the language abilities we want to measure from the effects of other factors, and this is a particularly complex problem" (Bachman, 1990, p. 161). Especially in speaking assessment, it is difficult to identify the source of the error because it is a matter of decision whether to consider the test taker's accuracy in language or his/her ability to cover the topic. Therefore, the raters or the teachers have to come to an agreement on the aspects they will consider beforehand so that they are clear about what to pay attention to while assigning scores.

2.1.3 Authenticity

Authenticity is the degree to which the features of the test are related to the features of language used in other contexts (Bachman, 1990). The two approaches to authenticity are the real-life approach and the interactional ability approach. The first refers to tests of language as it is used in real-life. This approach, however, has been considered naive as testing and real-life settings are not the same, that is to say no matter how close testers try to bring the testing context to the real life one, they will not succeed (Spolsky, 1985, cited in Sook 2003). The latter approach refers to language tests, the authenticity of which comes from their 'situational' and their 'interactional' authenticity. 'Situational authenticity' is the extent to which test characteristics are related to the ones occurring in different situations while using the language in target situations. 'Interactional authenticity' is the extent the test takers' language ability is involved in completing the test task (Bachman, 1991, cited in Alderson and Banerjee, 2002). Later, however, Bachman and Palmer (1996) refer to authenticity as a single quality of test usefulness. They separate it from the notion of 'interactivness' and define authenticity "as the degree of correspondence of the characteristics of a given language test task to the features of a TLU [target language use] task" (Bachman and Palmer, 1996, p. 23). They consider this aspect of a test important as it determines the "potential effect on test takers' perception of the test and, hence, on their performance" (Bachman and Palmer, 1996, p. 24). The authors believe that the relevance of the test tasks and language features may create a positive attitude towards the test and therefore may foster effective performance on the test. However, Lewkowicz (1997, 2000), citing the results of a number of studies, paints a somewhat different picture. The researcher states that according to one study, students who took different language tests showed more preference for the test which was more familiar but less authentic than for the test which addressed their needs in the use of the target language. She concluded that students seem to be more concerned with test familiarity and difficulty rather than test authenticity. Moreover, Lewkowicz found no difference in performance resulting from test authenticity (cited in Alderson and Banerjee, 2002).

2.1.4 Interactiveness

Interactiveness is a feature of a test which refers to the degree to which the test taker's 'individual characteristics', that is the 'language ability, topical knowledge and affective schemata', are considered in designing test tasks. Thus, the interactiveness of the test task is maintained if the aforementioned aspects of language use are addressed in the test task. For example, if "a test task requires a test taker to relate the topical content of the test input to her own topical knowledge it is likely to be relatively more interactive than one that does not"(Bachman and Palmer, 1996, p. 25). Therefore, unlike authenticity, where a relation needs to be created between the use of language in test task and that of real life, interactiveness refers to the relation between the test taker and the test task.

2.1.5 Impact

Another important issue to take into consideration while constructing language tests is the impact that the test may have on individual test takers and/or the educational system. One aspect of impact is 'washback' (Bachman and Palmer, 1996), which has been a subject of study and discussion for many researchers. Hughes (1989) describes backwash (the term he used for 'washback') as "the effect of testing on teaching and learning" (p. 1). Cohen (1994) refers to it in a broader sense, that is the way the "assessment instruments affect educational practices and beliefs" (p. 41, cited in Bachman and Palmer, 1996, p.30). Washback can be harmful and beneficial. If the content and the objective of a test are very different from those of the course, the test may have a harmful backwash effect. On the other hand, if the test and the course objectives are connected with each other and one is used for the improvement of the other, the test is said to have a beneficial backwash effect (Hughes, 1989, pp. 1-2). Teachers should test students on the skills they want students to learn and develop. For example, if teachers want their students to be able to communicate in different life situations successfully, they have to test them on those skills. This interrelation between the teaching and testing may lead to a positive backwash effect and to the development of a language teaching program, such that relevant changes may be made in the program based on the test requriments (Bostwick and Gakuen, 1995, cited in Sook, 2003).

Another factor that may determine the degree of impact of a test is the reason for taking the test and the importance of the test scores in students' lives, whether students are taking a high-stakes or a low-stakes test (Shohamy, 1993). Tests are considered to be high-stakes if they have a direct impact on students' immediate future, such as admission, promotion, placement; while low-stakes tests are those which do not have direct impact on students' immediate future, such as classroom tests (Davidson et al, 1997, cited in Clapham, 2000). However, Clapham (2000) argues this view stating that there should not be any difference in tests because "even if the results of a classroom test do not affect a student's immediate future, the results may become self fulfilling" (p. 151). In a classroom, for example, a low score may affect the student's future performance just because that student will be considered a poor learner both by the teacher and the students (Clapham, 2000). However, a study conducted in the Israeli educational system, which aimed to observe the impact of an Arabic Second Language Test (ASL) and an English Foreign Language Test (EFL) over a period of time, shows a different picture (Shohamy, 1993). The results of the study showed that the impact of the ASL test has decreased over time while the impact of the EFL test has increased resulting in many significant changes in the course, such as introducing new activities and new course materials. The reason the author mentions this result is that ASL was not a high-stakes test and the results of that test were not used in making important decisions. On the other hand the EFL test was a high-stakes test and, therefore, it had higher impact on the educational system (Shohamy, 1993).

In another study carried out in Hong Kong secondary schools with the aim of examining the washback effect of a newly introduced Hong Kong Certificate of Education Examination in English (HKCEE) exam on teaching and learning, showed that "among the different aspects of teaching and learning, teaching content has so far received the most intensive washback effect in Hong Kong secondary schools — thus the area of washback intensity" (Cheng, 1997, p. 51). The most salient change was observed in the content of teaching since after the exam was introduced most schools changed their textbooks. However, the author considers the data of the study not to be sufficient to make judgments about the effectiveness of the washback effect. The author also states that in this kind of study it is of paramount importance "to analyze the nature of the test type, be it a large-scale public examination or classroom assessment since the function and/or the stake a particular test bears determines the degree of its influence and the areas of its washback intensity" (Cheng, 1997, p. 51).

As the decisions made about students based on the scores they earn on the tests (no matter high- or low-stakes) have a direct impact on students, an important issue to take into consideration while assigning the scores to students is fairness. Fairness, as Farhady (1999) defines, "refers to a value judgment regarding decisions or actions taken as a result of test scores. It involves a comparison between the decisions which were made and the decisions which should have been made" (p. 36). Test makers should make sure that all test takers are treated equally (Bachman and Palmer, 1996). Kunnan explains fairness in the following way:

> The first fairness concern is whether test-score interpretations have equal validity for different test takers and test taker groups as defined by salient characteristics such as age, gender, race/ethnicity, native language and culture, physical or physiological impairment, and opportunity to learn. In other words, the concern here is primarily in the area of fairness of individual instruments and their score-interpretation. A second definition of fairness is about the concern that goes beyond the "equal validity" concern to the concern of social equity, such as access to higher education, employment, immigration, citizenship, certification or career advancement. In other words, this type of fairness is concerned with the fairness of how instruments and their score-interpretations work in society (Kunnan, 1999, p. 238).

Kunnan (1999) admits that much work needs to be done to establish the role of fairness in different aspects of language assessment, however, while Rawls (1971, cited in Kunnan, 1999) uses the words 'justice as fairness' to refer to central ethical theory, Kunnan (1999) reverses those words and has 'fairness as justice' to refer to the issue of fairness. "At least this much should be clear at this juncture. That fairness could lead to justice for test takers and test users alike and that this is certainly a worthy goal to pursue" (p. 239).

2.1.6 Practicality

An essential quality for a test is practicality. While the above-mentioned qualities refer to judgments gained through scores, practicality refers to ways the test will be constructed and implemented. (Bachman and Palmer, 1996). A test has to be practical which refers to issues like "economy, ease of administration, scoring and interpretation of results" (Bachman, 1990, p. 34). It is the teachers' responsibility to make the tests as short and practical as possible with the most efficient way to obtain information about students' knowledge because schools cannot afford having tests which need special equipment or highly trained examiners or raters (Weir, 1993).

All the above-mentioned qualities of assessment are general and have to be addressed in constructing any test which tests different skills of the language. To be able to relate them to oral tests, which is the point of the present study, we need to address some issues concerning the nature of speaking and the assessment of that skill.

2.2 Nature of oral performance

Many foreign language learners, if asked, find speaking harder than all other skills of language learning, that is reading, writing and listening (Nunan, 2003). One of the reasons, as the author points out, is that "unlike reading or writing, speaking happens in real time: usually the person you are talking to is waiting for you to speak right then." (p. 48). In contrast to writing, in speech you can neither edit nor revise what has already been said.

The nature of speaking has interested researchers for years. Referring to language acquisition research, Nunan (2003) points out that people do not learn language segments separately and then put them together to communicate, but they learn those segments by communicating in the target language. That is, learning/acquiring a second language happens the same way children acquire their first language. According to Shumin (2002), learners of a foreign language have to use the language in context in real communication, in order for them to be able to acquire the knowledge of how the language is used by native speakers. Here many 'factors interact' which make it difficult for language learners to speak fluently in a foreign language. Therefore, it is not of great use to spend time teaching speaking through bits and pieces of the language thinking that students will be able to connect them later on and start to speak. Real life conversations do not sound the same way as the 'conversations' of the lesson, which have the form of reciting dialogues (Nunan, 2003, pp. 49-50). Teaching speaking in a foreign language should be taken beyond the level of teaching rules (Shumin, 2002), as "language proficiency is not a unidimentional construct but a multifaceted modality, consisting of various levels of abilities and domains" (Carraquillo, 1994, p. 65, cited in Shumin, 2002, p. 206).

2.3 Assessment of oral performance

Given the fact that speaking is a complicated part of the language to acquire, it follows that its assessment will also be complex: "the sound of speech is a thorny issue for language assessment" (Luoma, 2004, p.10). Despite its complexity, with the development of communicative language teaching, the assessment of oral proficiency has gained more importance and attention (Nakamura 1993, cited in Sook, 2003).

Weir (1993) points out that real assessment of oral proficiency takes place if there is the requirement for the "candidates to demonstrate their ability to use language in ways which are characteristic of interactive speech" (p. 30). That is to say, the multiple choice tests, which are paper and pencil tests of spoken language, are of no use as they do not check real speaking ability. Weir (1993) states that:

> The important role of context as a determinant of communicative language ability is paramount. Language cannot be meaningful if it is devoid of context (linguistic discoursal and sociocultural). The context must be acceptable to the candidates as a suitable milieu for assessing

> their spoken language ability. The conditions under which tasks are normally performed should obtain as far as is possible in the test. A conscious effort should be made to build into tests as many real-life conditions as are feasible and considered worthwhile by the test writers and their peers (p. 30).

Taking into consideration the complicated nature of assessing spoken language, Weir (1993) suggests that certain criteria or guidelines should be developed for measuring test taker's ability to speak in a foreign language. These guidelines for assessment usually take the form of rating scales, the characteristics of which as well as their usefulness will be discussed in the following sections.

2.4 Rating scales

Research considering the testing of spoken language has a history of more than 20 years. A considerable number of studies have been carried out in the field of designing speaking tests in North America in 1980s, the focus of which has mainly been on content validity, construct validity, concurrent validity and reliability and rating procedures (Taylor, no date indicated). However, those earlier studies fail to address several important issues in testing spoken language. One of those issues is "the nature of rating instruments being used" (p. 1). Towards the end of the 1980s and beginning of the 1990s language test designers came to realize that different factors should be considered while constructing tests of spoken language and those factors need to be identified, developed, and explained according to the way they interact with each other.

For testers to be able to make relevant inferences from the assessment results, it is necessary for them to explain those results. Any score assigned to a student must be:

> ...put into context before it is meaningful. This is accomplished by comparing it with some desired state of affairs, goals or other information that you have that is relevant to your decision. Once this is done anticipation of the information is possible and finally a decision can be made about how to proceed (Genesee and Upshur, 1996, p. 36).

Luoma (2004) points out that while creating a rating scale for spoken language, it is of great importance to take into consideration "the special nature of spoken grammar and spoken vocabulary" (p. 27). Bachman (1990), however, emphasizes not only grammar but also the fact that it is important to define all the points of the rating scale so clearly and precisely that it is easy for the rater to define in which category the student's ability falls. A study conducted by Lumley and Qian (2001), which focused on the aspects of language that influence the final score on two language tests, has shown "that perception of grammatical accuracy seems to have the strongest influence on scores" (p. 94, cited in Alderson and Banerjee, 2002).

The issue of the raters' ability to accomplish the rating is also important. As McNamara (2000) points out:

> [the] rating given to a candidate is a reflection, not only of the quality of the performance, but of the qualities as a rater of the person who has judged it. The assumption in most rating schemes is that if the rating category labels are clear and explicit, and the rater is trained carefully to interpret them in accordance with the intentions of the test designers, and concentrates while doing the rating, then the rating process can be made objective (McNamara, 2000, p. 37).

This issue is vital as several studies show that even in cases where similar scores are assigned to test takers, the discourse performed by the same test takers is often qualitatively different. As Alderson and Banerjee (2002) point out, citing Douglas (1994); Pavlou (1997); and Meiron and Schick (2000), in this case there is the inference that something is wrong either with the rating criteria or their interpretation as made by the raters. Therefore, both the assessment criteria included in the rating scales of oral performance and the way the raters interpret them should be of primary concern for test developers and researchers as "the validity of interpretations of ability depends on the criteria used to rate performance" (Luemly and Qian, 2001, pp. 94-95, cited in Alderson and Banerjee, 2002). The point that scoring of oral proficiency can be considered valid and reliable only if good criteria for scoring are developed, the testers are trained to use them properly, more than one tester tests the same performance and "irrelevant features of performance are ignored" (Hughes, 1989, p. 110) is a way of summarizing the need for having good rating scales. However, not all the points mentioned by Hughes can be addressed in the classroom context because of the fact that there is only one person, the teacher, who is responsible for scoring students' performance. This is because classroom tests are considered to be low-stakes tests and lower reliability for these tests may be also acceptable, while in the case of high-stakes tests the level of reliability has to be as high as possible (Bachman and Palmer, 1996).

Luoma (2004) considers that a very important issue for the developers of speaking assessment instruments is that they have "a clear understanding of what speaking is like and then:

- Define the kind of speaking they want to test in a particular context;
- Develop tasks and rating criteria that test this;
- Inform the examinees about what they test;
- And make sure that the testing and rating processes actually follow the stated plans" (Luoma, 2004, p.28).

Bachman and Palmer (1996) approach the development of rating scales following two principles: the scale "may be either theory-based or syllabus-based" (p. 211). The different levels of rating scales which are used for proficiency testing and may be due to changes based on test purpose, are also categorized by addressing certain areas of language ability "with the lowest level in our rating scales defined as 'no evidence of the ability' and the highest level as 'evidence of mastery of' the ability" (Bachman and Palmer, 1996, p. 211).

It is important for testers to decide in advance whether the criteria of the rating scale will appear in a global/holistic form, that is language ability taken as a whole or in an analytic form, that is in a form where language aspects are addressed separately. The decision on whether to use the one or the other depends on "the degree to which one can describe in behavioral terms the different levels of proficiency that student performance will result in" (Weir, 1993, p. 45).

2.4.1 Holistic rating scales

Holistic scales, as Bachman and Palmer (1996) describe, are "one traditional approach to developing rating scales of language proficiency based on the view that language ability is a single unitary ability, and yields a single score, called a 'global' rating. Many such scales, however, contain multiple 'hidden' components of language ability" (p. 211). The existence of those 'hidden components' is probably the reason that the use of analytical rating scales is preferable in many cases. As "our understanding of the continuum of proficiency in speaking is currently limited" (Weir, 1993, p. 45), it would be better to use an analytic scale where each aspect of language is addressed separately. The reason the author has this belief is that there is no evidence that all students develop at the same rate in all aspects of language. Holistic scales, however, "collapse criteria together and assume that students progress equally in all criteria as they move up the band scale" (Weir, 1993, p. 45). An individual may gain the mastery of linguistic features faster than communicative competence and vice versa (Taylor, no date provided). An analytic rating scale seems to be more appropriate for classroom use as students are taught different aspects of language, and it would be better to test them on what they are taught addressing each aspect separately rather than assessing them based on one's overall impression. However, some international exams, such as National Certificate, American Council for the Teaching of Foreign Languages (ACTFL) and Test of Spoken English (TSE) are rated using holistic scales. Despite the suggestion of the use of analytic rating scales in assessing speaking, Taylor (no date indicated) does not categorically give preference to one or the other type of rating scale but just states that the developers of Speaking Tests in UCLES (University of Cambridge Local Examinations Syndicate) try to use the relevant points from both assessment tools.

2.4.2 Analytic rating scales

Analytic rating scales address different aspects of language ability separately.

> In situations where it might be useful to provide a single score, we recommend deriving this by combining componential scores which are arrived at by the use of analytic rating scales, rather than developing a single global rating scale. This is because of the problems involved with the use of global scales (Bachman and Palmer, 1996, p. 211).

A study conducted by Venema (2002) at the Japanese Technical University, which addresses the issues to be taken into consideration while constructing rating scales for classroom use, shows that the use of an analytic rating scale in a classroom context is advantageous as it gives teachers the opportunity to "become more explicit about performance criteria" (p. 4). The researcher believes that the process of construction of rating scales may be considered successful if it includes the grading rationale in them and also refers to students when necessary to "account for elicited performance" (p. 1). Bachman and Palmer (1996) bring up the same issue of including relevant guidelines in rating scales, suggesting that scales are defined "operationally in terms of criterion levels of ability, rather than as performance relative to other test takers or to native speakers of the language" (p. 212). The authors consider criterion-referenced scales to be advantageous for learners as they state the learner's level of ability as compared to certain criteria rather than to other people in the class or the native speakers of the language.

Another way of constructing a rating scale for assessing oral performance is described in the study conducted by McNamara (2001) in Melbourne. Considering the new decision about including speaking in the TOEFL test, the author has suggested constructing a rating scale based on the points the raters mention rather than using a standard rating scale. The oral performances from the piloted TOEFL test were presented to several raters who "were asked to carry out think aloud protocols as they listened to the performance" (McNamara, 2001, p. 3). Then based on what the raters noticed while listening to the performances, and based on the most salient points put down by them, further suggestions for the construction of a rating scale were made. Though, there is no evidence in the study which showed how the system worked in the case of assessment of speaking in the TOEFL exam, this may be a good way of constructing classroom rating scales as well because teachers know what aspects of language they have to pay attention to and they also know what they have to test the students on.

The development of classroom-based assessment scales is very important and is the point of this study. Therefore, the next two sections will cover some of the ways of constructing rating scales for classroom use and relate this to the situation in Armenia.

2.5 Rating scales for classroom use

The rating scales discussed in the previous section assess students' proficiency level in the target language, and they are mostly used in high-stakes tests. Classroom tests, on the other hand, are low-stakes achievement/progress tests; therefore, the assessment tool needs to be based on the curriculum, that is what the students cover during a certain period of time. In this regard, the existing rating scales are not appropriate for classroom use as in the classroom context the situation may be different, as teachers should not test students on something that they have not taught them. Both analytic and holistic scales assess the test taker's proficiency in language and range from zero knowledge level to native-like speakers and these levels may rarely exist in classrooms. However, it is possible to develop a new tool of assessment using some issues addressed in the rating scales and some issues specific to the particular case/classroom. This will also create positive washback for students, because, if they are tested on something that is not new or unknown to them, they may be motivated to learn better.

A precedent for this kind of work is provided by Lewkowicz and Nunan (2004), who have developed a special rating sheet for assessing both students' writing and speaking performance. The purpose of the work was to develop an assessment package for teacher training for Hong Kong secondary school teachers. That rating sheet provides information about what the students are expected to learn and know during the year, called "general criteria for assessing speaking/writing", and information about what the students are expected to know for a specific lesson, called "task specific criteria for assessing speaking/writing" (Appendix A).

2.6 The case in Armenia

Assessment in Armenian secondary schools is carried out by using scores ranging from one to five. This assessment tool is used to assess students' everyday performance in different skills. So if a student gets a 'five' on a particular day it means that different factors, such as home assignment and in class participation are considered while assigning the score. The average of the scores gained in everyday lessons and also the scores gained from the tests is the final end-of-year grade of each student. Given the assessment situation in Armenia, the definition of the term 'assessment' and for the purpose of the present study, the term 'assessment' will be used throughout the study.

Few studies have been conducted in Armenia concerning methods of classroom assessment; however, there are several studies describing the current situation and suggesting what needs to be considered and improved.

Harutyunyan (2002) conducted a study, the aim of which was to study what the situation was in the sphere of classroom assessment in Armenia. The study included secondary schools from all the regions of the country and the capital Yerevan. The study included both teachers and the students of different schools. The findings of the study were based on two questionnaires, one of which was for teachers and the other for students. The researcher found that English teachers in Armenia base their regular classroom assessment on students' oral performance, "among the activities graded by the teachers, oral answers are the most common (95%)" (p. 62). The majority of the students, 85%, confirmed this by choosing the oral answers as the most frequently occurring technique of classroom assessment. This is worth consideration for two reasons according to the researcher. First, basing the scores mainly on oral performance, teachers do not consider the development of other skills of language (reading, writing and listening). Second, they assess students in the absolute absence of criteria. The same research presents the finding that 82% of the respondents, that is the teachers of secondary schools in Armenia, have stated that they base their assessment on their teaching experience and only 26% base their assessments on criteria "agreed upon with colleagues", while 22% use criteria "offered by school or local authorities". The researcher expresses her concerns about this, as it "cannot be regarded as a reliable basis for objective assessment. Only state standards backed up by systematic training and local standardization can ensure reliable assessment in this area" (p. 63).

The same concern is expressed in another study conducted in Armenia by Muradyan (1999). She considers the lack of descriptors of scores to be the reason for subjective assessment. The following description shows the real situation:

> The problem is that by leaving oral assessment to the teacher's discretion, every teacher's individual perception of the goal and purposes of assessment and ways of achieving them would be of paramount importance. Very often what one teacher would consider important, another would consider unimportant. Such categories as effort, participation or even attendance would affect one teacher's grading while having little impact on another's assessment. As there were no scoring descriptors for assessing oral answers, nobody at least officially cared about that aspect of marking. Different teachers used to have different reputations: 'tough grader' or 'lenient grader' labels were far too well-known to me: it was basically every teacher's commonsense experience and/or teaching philosophy that guided him or her in deciding the mark of a student (Muradyan, 1999, p. 24).

Muradyan further states that she does not think that in the case of having guidelines absolute objectivity will be guaranteed but at least the subjectivity will decrease in assessing students' oral performance. Therefore, she argues for greater transparency and fairness of assessment of oral performance.

Given the complicated nature of spoken language, the importance of its accurate assessment and the above-mentioned evidence, which shows that Armenian secondary schools lack methods for assessing students' oral performance properly, the present study intends to find a solution to the existing problem by suggesting a rating scale which includes the aspects of language that are taken into consideration while assessing oral performance. There is the assumption that the use of this newly-developed rating scale, which is the primary objective of this study, will provide an opportunity for teachers to assess their students more objectively.

Chapter Three: Methodology

3.1 A Descriptive study

This study has been conducted using descriptive research methodology. Descriptive study, as defined by Farhady (1995), gives opportunity to the researcher "to describe and interpret the current status of phenomena" (p. 144). To get a more thorough understanding of some of the characteristics of the phenomenon of speaking and the effectiveness of its assessment, a case study was conducted. Farhady (1995) notes that "in a case study, a researcher makes an intensive investigation of a social unit" (p.149). He further suggests that in such studies the researcher investigates the past and present situation of the sphere of concern and other factors which contribute to the state in which the field under investigation is. This kind of study helps the researcher to end up with a "comprehensive description and clear picture of the unit under investigation" (Farhady, 1995, p. 149).

3.2 The process of the data collection

The aim of this study was to develop an appropriate rating scale with different categories and descriptors for each grade and determine its effectiveness in assessing students' oral performance in the fifth grade of Armenian Secondary Schools. Therefore, this chapter will cover the process of the development of the rating scale, the selection of the participants, the process of using the developed rating scale in the classrooms and the feedback received from the teachers who used the rating scale in their actual classrooms.

3.2.1 The development of the rating scale

Given the fact that in Armenian secondary schools the assessment of students' knowledge of English language is mainly based on their oral performance (95%) and it is done in the absolute absence of any criteria (Harutyunyan, 2002), the need for the development of some kind of a rating scale seemed to be a priority. Standardized rating scales are not appropriate for classroom use for two major reasons. First, those scales are usually used to assess students' proficiency of the language and to establish the level of the knowledge that students have. This is not the case in the classroom as in the classroom students are tested on what they are taught, that is, their achievement in the language. Second, standardized rating scales range from zero knowledge to near native proficiency, which also does not happen in the classroom. A new rating scale was developed on the basis of the existing rating scales which are used to assess oral and written performance in international and well-known exams and also considering the aspects of the language that the students cover at a given level.. Though, some categories given in those rating scales have been addressed in the newly formed rating scale, relevant changes have been made taking into consideration the level and the environment where the scale is to be used and implemented.

In the development of the new rating scale, three main points were considered. First, the criteria and descriptors of the scores were based on the curriculum and the textbook the students of the fifth grade study, that is, testing the students on what they have been taught. Second, all possible spheres of oral performance that are applicable at a certain level, that is grammar, vocabulary, content, fluency, communicative strategies and comprehension were covered. Third, the issue of making it easy and practical for teachers to use and be able to make relevant changes, if necessary, was considered.

The work done by Lewkowicz and Nunan (2004) was used as an example, as a precedent study in the same sphere (Appendix A). In this rating sheet there are two sections, one section covers all the categories and the expectation from the students regarding what they have to know, based on the curriculum during one educational year. The other section covers the categories and the expectations of what students have to know on the particular day. In the second section the scores for each category are also included.

Presently, in secondary schools the scoring system ranges from one to five. Therefore, not to make fundamental changes in the grading system and not to confuse teachers, the same five-scale range is kept in the newly developed rating sheet. Only point one is viewed differently in this rating sheet, because despite the fact that this point exists in the presently used scoring system, it is on rare occasions, used only with the purpose of punishing students. In the developed rating scale one is just the lowest level of knowledge (see Appendix B). However, this five-scale scoring is used separately for each category (for grammar, vocabulary, content, fluency, etc.) and the final score of the student is the mean score obtained form these different categories. To help teachers differentiate what each score means, the description of each score was also provided on the rating sheet (1 – poor (below minimal expected level), 2 – needs improvement, 3 – satisfactory, 4 – good, 5 – very good). There was no assumption that by providing different categories and descriptors for the grades absolute objectivity could be achieved in classroom assessment but at least the subjectivity would decrease. That is to say, in any case the assessor/teacher may have his/her personal approach to the one, two or five, no matter how the range of the scores are described; however, when there are different categories and guidelines, the teachers may be more careful about what they do. What is more important, the assigned scores may be more reliable and fair. These are of primary concern for any assessment process.

3.2.2 The selection of the participants

The aim of this research was to develop a sample rating sheet for assessment of oral performance which will be subject to changes to suit a relevant level. The selection of the fifth grade students was connected with the fact that the teachers who volunteered to participate in the study taught classes at that level. Therefore, two teachers from two different schools and their students (six classes, about 90 students) of the fifth grade participated in the study.

3.2.3 Procedure

Two English lessons were recorded in the above-mentioned classes to establish the level of oral performance in the fifth grade and to set examples for the description of each score provided in the rating scale. Fragments of the recorded lessons were transcribed (Appendix C) for explanation of the different scores that were used in the rating scale.

The teachers used the rating scale to assess the students' oral performance in their actual lessons for a month. Before using the scale in the particular lesson, relevant changes were made in the rating scale, especially in the grammar category, based on the daily task and also some considerations that the teachers expressed during the discussions that we had after each lesson. The reason the main changes were made only in the grammar category was that at this level the main focus is on this aspect of language and the tasks and requirements change each day in this regard.

Grammar	**Grammar** 1 2 3 4 5
is accurate in using • tenses (different forms of present and past) • conditionals (first and second) • passive voice • constructions with **used to** • reported speech	• is able to use and identify the tense forms (past tense for speaking about past events and present for general events) • is able to switch from one structure to the other appropriately

The column on the left of the fragment of the rating scale, for example, includes criteria which students are supposed to master by the end of the academic year. The column on the right represents criteria which students are supposed to know on a particular day as well as the scores which the teachers had to assign to students considering their proficiency of performance on each mentioned point. The above fragment of the grammar section includes the primary description for the fifth level students. However, later after a careful observation of both the book of that level and also the daily tasks, relevant changes were made in both columns. For a lesson which focused on the correct use of tense forms relevant points were included in the grammar category which is the case in the above fragment. While for the next lesson, when the teacher had explained the use of reported speech, the focus of the lesson had to be on how well students mastered it and whether they could use it appropriately, the following points were included in the grammar category of the rating scale:

Grammar	**Grammar 1 2 3 4 5**
is accurate in using • tenses (different forms of present and past) • conditionals (first and second) • adjectives (the degrees of comparison) • nouns (plural forms, countable and uncountable) • passive voice • constructions with **used to** • reported speech	is able to • use and identify reported speech • switch from direct speech to indirect and vice versa freely • identify the difference between the past simple and past continuous • switch from one structure to the other appropriately

Teachers were also asked to inform the students about the fact that certain criteria are used to assess their knowledge and different categories are included in the rating process. That is, the teachers showed the students the new assessment tool and explained all the criteria that would be considered while assessing their performance. The purpose was to have a chance to observe also the students' reaction and attitude to a new way of assessment which was different form the one they were used to having and let them know what they were assessed on.

After using the rating scale for some time, the teachers were invited for an interview to share their opinion about the usefulness of the rating scale. The interview was a semi-structured interview. The reason for choosing this type of interview was to avoid the restriction of the answers from the teachers' side and the freedom of discussion of points not relevant to the study. The advantage of this kind of interview is that "the level of the interview can be adjusted to suit the situation and the interviewee" (Farhady, 1995, p. 221). Eight basic questions were prepared for the interview. The two teachers were interviewed separately and were asked the same questions. However, based on the way the discussion developed, some additional questions were asked of each of the teachers. This is the reason that there is a difference in the number of the questions in the transcript of the interviews. The language of the interview was English. The interviews were recorded and transcribed (Appendix D and Appendix E). The questions and answers to the questions will be discussed in the next chapter.

Chapter Four: Results and Discussion

This chapter covers the findings from the observations of the lessons and the interviews, which were conducted as described in the previous chapter.

4.1 The observations

The first stage of the study was observation of the lessons of the fifth level. The aim of the observations was to establish the type of tasks students had to complete during the lessons, which helped later in determining the relevant categories in the rating scale. It was obvious both from the tasks included in the book and the activities held during the lessons that the assessed oral tasks were mainly role plays, retelling of texts, and sharing experiences with each other. Depending on the topic of the lesson and the aim of the teacher, different activities were conducted during the lessons. For example, if the lesson focused on a story, the oral activity of the lesson would be retelling the text with a follow up discussion. If the lesson was based on students' lives and their experiences, another type of discussion would take place where students shared their experiences asked and answered both teacher and peer questions. Taking into consideration that in such activities certain aspects of language should be considered, relevant categories were included in the scale, that is, grammar, vocabulary, fluency, content, communicative strategies and comprehension.

At this stage most activities were guided by the teacher, which is good because students in some cases try to concentrate on the topic that is the content of what they are talking about and the accuracy of the speech suffers. In other cases they concentrate on the accuracy of the speech and the fluency and the communicative strategies suffer. Therefore, some teacher guidance was needed to maintain the balance between those aspects. The guidance that the teachers had took the form of asking questions or adding some information the students obviously wanted to say but had some difficulties in doing so. The teachers corrected the grammar mistakes in the following way: when students made obvious grammar mistakes, the teachers repeated the sentence with a correct form rather than stopping the student, correcting the mistake and explaining why s/he could not say the sentence that way. This was done (as the teachers reported in later discussion) to encourage the students to speak and gain fluency. If they were stopped for each mistake, they would not want to speak even if they had something interesting to say.

The second stage of the study and the observations was recording two actual lessons to have evidence that could show quality of speech for points of each score that would be included in the rating scale. The assessment of the oral speech is based both on accuracy and fluency with much attention on fluency as one of the teachers asserted later during the interview (Appendix E, TB, Q8). That is to say, some minor grammatical mistakes do not determine the score assigned to the students; the obvious mistakes are considered while assigning the score. Thus, for example, in the following extract, where regardless of some minor language problems and also use of the mother tongue, there is no communication breakdown; the student follows the point of the questions and gives relevant answers, this is considered a 'very good' answer, which is a 'five', the highest score in the scale.

T: Now speak about your family.
S2: Family, ok. We live here I, my mother and father. I have aunts and uncles. Some are in America and the other part is in Lebanon.
T: Have you ever been in America?
S2: No.
T: Are you going to visit your uncles in America?
S2: I don't know but, er..
T: Ah, they are visiting you and you don't need to go there.
S2: Yes!
T: Do you like your father?
S2: Sure...
T: Is your father very serious?
S2: My father? My father is not very serious 'artakinits' (appearance) how can we say.
T: Appearance.
S2: Yes.
S3: Because he has mustache and er....

However, in cases when students make mistakes which they are not supposed to, that is incorrect use of the verb 'to be', verb tenses present, past and future, they get a lower grade.

For example, in the following extract when the student uttered the sentence "I very like Moscow" even the whole class reacted with laughter and repeated the sentence in a way which showed that at that level they should not make such mistakes.

***S3:** I born in Russia,*
***T:** I was born*
***S3:** I very like Moscow.*
***Class:** I very like (other students repeat and laugh to show that their friend made a mistake)*

Therefore, in such cases, the score assigned to the student is 'two', the explanation of which is 'needs improvement' in the rating scale or a 'three' (depending on further development of the speech), the explanation of which is 'satisfactory' in the rating scale. In the following extract there is a little communication breakdown as the student generally says what she wants to say regardless of the questions. She also avoids speaking about topics she does not like.

***S4:** I born in Yerevan.*
***T:** I was born ...*
***S4:** I was born in Yerevan, in 1993. I used to play with those house. My favorite food is pizza, only pizza, I like*
***T:** Who is your favorite hero?*
***S4:** I like every one and I go to school, in school my favorite subject is Mathematics, Armenian, English, Russian.*
***T:** Have you ever been to Russia?*
***S4:** Yes, once.*
***T:** Speak about that.*
***S4:** I don't remember because I was very small.*

This speech is considered of average performance because the student avoids answering some specific questions, yet she speaks about what the class discussed during that particular lesson and the meaning of her speech is quite clear.

The teacher is the person who makes the decision as to what score to assign to the students even in the case of the existence of categories, guidelines and explanations of scores in the rating scale Therefore, there are two major concerns at this point. First, it is very important that the teachers are clear about how to conduct the assessment. Second, they should try to get rid of their initial perception of the 'one' and 'two' as failure scores and make use of them as explained in the rating scale.

The next section covers the results of the interviews with the teachers, which will help give a more comprehensive picture of their perception and interpretation of the rating scale. The aim of the interview was to get feedback from the teachers after they had used the rating scale in their classrooms.

4.2 The interviews

Eight basic questions were prepared beforehand and some additional questions were asked during the discussion. The questions and the answers will be addressed in relevant sections referring to the (1) advantages of using the rating scale, (2) difficulties of using the rating scale, (3) students' reaction to using the rating scale, (4) recommendation for changes to the rating scale and (5) summary of the interviews and the findings. A discussion concerning some issues on the distribution of the scores in the rating scale and the implementation of the rating scale will also be addressed in this section.

4.2.1 Advantages of the rating scale

Though at first the teachers encountered some difficulties while using the rating scale, both of them found that using the rating scale in classrooms for everyday assessment was useful and advantageous. The main positive points that they addressed were: a) it helps reveal student performance in different aspects of the language; b) it motivates students to learn better to achieve a specific level; c) it is a way of teaching new aspects of language; d) it helps to avoid cases of complaints concerning the subjectivity of assessment.

4.2.1.1 Range of students' performance

The teachers considered the rating scale to be very useful as it raises both the teachers' and the students' awareness of the aspects considered in the assessment process. Though the teachers stated that before becoming involved in the study they were used to assessing students' overall performance and found that less time consuming than using this rating scale, they stated that it is better to use this instrument of assessment as it fosters learning and effective assessment. It was clear from the teachers' answers to the following questions "Do you think this is a good tool of assessment for everyday use?" and "Was there any difference in the scores students got from the previous way of assessment and this new assessment sheet?" (Appendix D, TA, Q1); "What do you think are the advantages of using this means of assessment?" (Appendix E, TB, Q6).

4.2.1.2 Motivates students to learn

The two teachers confirmed that it was difficult for students to get used to using such an assessment tool in the classroom. However, when the teachers explained what each category meant and how each of them would be addressed while assessing the students, they tried hard to learn how to get the desired score, that is, as high as possible. Though usually students try to get the highest score which is five on the rating scale, there are students who were aware of their limitations and that four or three was their desired score.

> They were excited to know that they were graded in a different way. They asked "what is this?" and I explained what categories were included there. They got motivated to perform accordingly to get higher grades. (Appendix E, TB, Q4)

Even weak students got involved in the lesson and started to participate to score higher, as Teacher A indicated:

> There was a case when a girl, who was very reserved and was not active during the lesson, was motivated by it and volunteered to participate and show that she also can. This helped her to be involved in the lesson and as she liked this process and this way of assessment, she came to the class prepared so that she may perform accordingly to get higher scores (Appendix D, TA, Q8)

4.2.1.3 Teaches new strategies

Apart from motivating students, the rating scale created opportunities to teach students new strategies. For example, the category "communicative strategies" turned out to be something new for both teachers and students, and it was strange for students that they could be assessed on this aspect of language. Realizing the importance of communicative strategies in communication in the foreign language, the teacher decided to teach students how to address those strategies while retelling a text or while talking to their peers or the teacher.

4.2.1.4 Reduces subjectivity

The rating scale helps the teachers be less subjective in their judgments about the students' performances. Now they had an explanation for the score assigned to a student. This is a way that avoids students' complaints about the scores they get. To the question "What do you think are the advantages of using this means of assessment?" Teacher A gave the following answer:

> This rating sheet helps the teacher answer the students' questions/students' "why's" after having their grade. Usually the students ask "why a four?" "Why a three?" Now they know that they deserve certain grade because of several reasons. While student(s) speak both me and the rest of the students note down the mistakes that the person makes without interrupting the students and then we have the chance to both discuss the mistakes and also give reason for a certain grade. (Appendix D, TA, Q9)

Teacher B's answer to the same question did not refer to scoring directly, but from what she said, it can be inferred that she was referring to reducing subjectivity:

> ...students are more conscious of the way they are assessed which means that they are willing to study better. It fosters students' learning as well because if they know that they are going to be assessed according to their knowledge of grammar, vocabulary and other skills they prepare more carefully (Appendix E, TB, Q6).

If students are conscious of what and how to study, they feel more responsible for the score they get because they are aware of the demands for each performance.

4.2.2 Difficulties in using the rating scale

The teachers identified two main difficulties while using the rating scale. The first one was the fact that using this assessment tool is more time-consuming than the manner of assessment they were used to using before. The other was connected with the existence of categories that the students are not familiar with and are not used to using.

4.2.2.1 Time

The teachers stated that before having this assessment method they based their judgments on their overall impression, but in this case where they have to address so many aspects, time is of great concern. However, for Teacher B this was of more concern than for Teacher A, as the latter found an economic way of using the scale by writing the categories on the board and also asking students to help her in assessing their peers. Teacher B, however, noted that it took her some time to assess each student every day. However, she believed that it was only matter of getting used to using the scale. If it were to become an everyday process, it would not take that much time.

> I think it is a good one. The only matter is that it is a bit time consuming. It took me three to five minutes to assess each student every day. However, I think if one gets used to it, it will be easier to use and consequently it will be more effective (Appendix E, TB, Q1)

4.2.2.2 Communicative strategies

Among the difficulties the existence of the category "communicative strategies" was also mentioned. When students talk, they try to concentrate on the topic they are going to speak about and they do not bother to think where they look and whether their interlocutor follows them or not. "The students are used to stand and look at the walls while they tell something" (Appendix D, TA, Q2). In this regard, assessing students on this category was challenging for Teacher A. She stated that she had to, first, explain what it was and its importance, and then, assess how successfully they used those strategies.

4.2.3 Student reaction to using the rating scale

Despite the fact that students first were surprised and not happy to see a category in the rating scale that was not of much concern for them before, that is communicative strategies, the students seemed to like this way of assessment. As the examples of the previous section and the teachers' answers show, students were really excited with the new way of assessment when they understood what each category meant and realized their importance. The reason the students were excited and happy was probably not only connected with the fact that they liked the categories, but the fact that they were informed about the way they were being assessed and they had the chance to participate in that process. In this regard, Teacher A told of an incident which reflects the students' attitude towards the new way of the assessment:

> I wrote the .categories on the board during the break and so did not spend class time on it. Later, which surprised me very much; my students wrote the categories on the board before the lesson. They got really involved in the process and showed more interest both towards learning and the way they were assessed (Appendix D, TA, Q5).

4.2.4 Recommendations for changes to rating scale

Though reluctant to suggest any changes in the developed rating scale, the teachers mentioned some aspects which they would like to see modified. The two most important aspects that they mentioned were the range of scores provided in the rating scale and some of the categories. Each will be discussed in turn.

4.2.4.1 Range of scores

Teacher A noted that seeing the rating scale first she had thought that "it would be better to have more scores in the range, but given the fact that the usual assessment method is kept, it is ok" (Appendix D, TA, Q12). She mentioned exactly the same reason that the researcher had for keeping the scoring range the way it existed, while developing the rating scale. It was done for several reasons. First, as it is difficult for teachers to get used to a new way of assessment, it would create additional problems for them to have a scoring range of, for example, from one to ten. They already have a preconception of the scores from one to five and it is easier for them to identify where the students stand using these. Second, it would interfere with the whole grading system of the secondary schools as well, because the range 1 – 5 is used for assessing all other skills of English language (also all other subjects), having a different one for only speaking, would confuse both the teachers and the students.

4.2.4.2 Categories

As Teacher A mentioned the difficulties that her students had with communicative strategies, she was asked whether she would like to take that category out of the rating scale. She refused categorically and said, "I think it is a step forward" (Appendix D, TA, Q13). She agreed that it was difficult for students to learn them but as she considered communicative strategies very important for successful communication, she said she would better have them in the rating scale and teach students to communicate accordingly. Moreover, the existence of that category in the scale and the fact that it contributes to their final grade, helped motivate the students to use such strategies.

While speaking about suggested changes in the rating scale, the only concern that Teacher B had was the number of categories included in it. She suggested reducing the number of categories. She was concerned especially about the grammar category. Though she considered grammar a very important aspect of language, she thought at level five more focus should be on fluency. As it is difficult for students to master fluency in speech, the teacher wanted to pay more attention to fluency of speech than accuracy. The stress on grammar, on the other hand, makes the students' grades suffer and therefore may demotivate students to speak;

> …at this level there is the issue of getting the students' fluency in communication and acquiring fluency is more important than accuracy as they know the grammar just while speaking they misuse it and their grades suffer from it. (Appendix E, TB, Q8)

Her point here was not to take the grammar category out altogether but try to reduce its weighting to encourage students to speak as freely as possible.

However, given the fact that both teachers claim that the students showed more interest towards learning because they knew how they were assessed and what was considered, one can assume that the grammar category is worth keeping.

4.2.5 Summary of interviews and the findings

Based on the above discussion it can be inferred that using this kind of a rating scale for everyday assessment in Armenian secondary school classrooms may be considered effective in several ways:

1. It addresses more aspects of the language and therefore the assigned score is more informative than the one assigned in previous way of assessment.
2. It may serve as a way to motivate students to learn better because they know that there are certain criteria and if they reach those criteria, they will score higher than they did before.
3. It helps teachers teach the students new aspects of oral communication, which are very important, for example communicative strategies.
4. It includes also students in the assessment process and thus reduces the possibility for any doubts about the fairness of the assigned score.
5. It helps the teacher give reasons and explanations for assigning certain score as s/he has the assessment sheets as a record of the student's performance.

4.3 Distribution of Scores

Additional information about the use of the rating scale can be observed by studying the way the scores were assigned to the students according to the rating scale. Table 1, on page 40, depicts the way the teachers assigned scores to their students for different categories. The rating of 21 students' oral performance is presented in the table. The students' names are not included in the table for protecting their privacy. The categories of the rating scale and the average score for each student are included on the left side of the column. On the right side of the column different scores assigned to students are included. This will help give a more vivid picture of the categories to see which one has the most impact on the average grade.

Here it is obvious that the teachers differentiate among the different skills of individual students. This is very important as the aim of using the rating scale in classroom assessment was to help teachers address different aspects of the language. There are examples, (the bold columns in the table), which show that this really worked. For example, student 21 scored 'two' for grammar; 'five' for vocabulary; 'three' for fluency and content; and 'four' for communicative strategies and comprehension. This shows that all aspects have really been considered and also the final grade does not suffer because of one or the other category, grammar in this particular case. The final grade for the day is the average score obtained from different categories. Teachers add up the scores obtained for each category and divide the number by six, the number of categories. There may be cases when the score is not a round number. In the case that the average score was 3.3, it would be rounded down 3 or if the average score is 3.5 or higher 3.6, 3.7 the total score would be rounded up to 4. This is the reason that all scores are whole numbers.

To see the students' range of scores for the different categories, it is worth having a look at Table 2, which depicts the number of students who were awarded 'one's, 'two's, 'three's, 'four's and 'five's for different categories. It is evident from the table that there are no cases when 'one' is assigned to a student in any of the categories of the scale. In the case of the 'two' four students were awarded this grade for grammar, one student for fluency and one student for communicative strategies. The number of students getting 'three', 'four' and 'five' seems to be similar. There are two possible reasons for this distribution of the scores: first, teachers may have used their initial perception of 'one' and 'two' as grades of

Table 1: Scores assigned to students for each category.

Student #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Grammar	3	4	3	5	3	3	5	4	4	4	5	3	2	2	3	5	5	4	2	4	2
Vocabulary	3	4	3	5	3	3	5	4	4	4	5	4	3	5	5	5	5	5	4	3	5
Fluency	3	4	3	4	2	4	4	5	5	5	4	4	4	5	4	5	5	5	3	4	3
Content	3	5	3	5	3	3	5	5	5	5	5	3	4	5	5	5	5	5	4	4	3
Communicative strategies	4	3	4	5	2	4	3	4	4	3	3	5	4	4	5	5	5	5	4	4	4
Comprehension	4	5	4	5	3	4	5	5	5	4	5	5	4	4	5	5	5	5	4	5	4
Average score	*3*	*4*	*3*	*5*	*3*	*4*	*5*	*5*	*5*	*4*	*5*	*4*	*4*	*4*	*5*	*5*	*5*	*5*	*4*	*4*	*4*

punishment and failure and tried to avoid assigning them to students and kept the main range of the scores from 'three' to 'five'. Second, students may really have performed well on their assignments. Classroom tests are achievement tests and students are assessed on what they are taught, therefore, they are expected to perform well.

Table 2: the number of students who got different scores for each category.

	1	*2*	*3*	*4*	*5*
Grammar	–	**4**	**6**	**6**	**5**
Vocabulary	–	–	**6**	**6**	**9**
Fluency	–	**1**	**4**	**9**	**7**
Content	–	–	**6**	**3**	**12**
Communicative strategies	–	**1**	**4**	**10**	**6**
Comprehension	–	–	**1**	**8**	**12**

Table 2 helps also in identifying how much emphasis/weight is given to different categories. If we study the categories and the number of students who got the lowest and the highest scores, it will be clear that grammar is the category which got the most weight. There are four students (the largest number) who were awarded a 'two' for grammar and five students (the smallest number) who were awarded a 'five' for grammar. The reason is probably that still teachers tend to focus on the grammar and the accuracy of the speech no matter that they assert that at level five more attention is paid to the fluency of the students. This supports the study conducted by Lumely and Qian (2001) referred to earlier in this research, which focused on the aspects of language that influence the final score on two language tests and showed "that perception of grammatical accuracy seems to have the strongest influence on scores" (p. 94, cited in Alderson and Banerjee, 2002).

The other two categories, where a small number of students were awarded the highest score and some were awarded the score of 'two', are fluency and communicative strategies. The idea of having communicative strategies as a category to be assessed while assessing speaking was new for the students as one of the teachers mentioned during the interview (Appendix D, TA, Q2). Therefore, students may have had problems using those strategies in the initial process of the use of the rating scale and got lower grades on that category. Fluency is the other crucial point in speaking and is given much attention by teachers; however, it seems that many students have not mastered it yet at this stage. The distribution of the number of the students in the rest of the categories seems to be more or less similar. The reason may be that students do not have many problems in dealing with those aspects of language.

4.4 Implementation of the rating scale

Considering all the above-mentioned points concerning the characteristics and the use of the rating scale, teacher training may be the first step needed to implement the rating scale in Armenian secondary schools for two major reasons. First, teacher training will help avoid possible difficulties the teachers may encounter while using the rating scale. Second, teacher training will help avoid possible negative consequences that may occur because of the misuse of the rating scale.

4.4.1 Training for the teachers

The results gained from the interviews showed that the use of the newly developed rating scale is effective for use in everyday classroom assessment. However, teachers may encounter some difficulties when left alone with the skeleton rating scale and are asked to use it to assess students. It is obvious that teachers need to be trained to be able to modify and implement the rating scale in their classrooms. As Nunan (2004) points out "almost any teaching task can be used for assessment purposes, and vice versa. The key difference is how the task fits into an instructional cycle and, crucially, what is done with the learner output from the task" (p. 143). Therefore, if a well-developed assessment tool is used in the classroom, teachers may find it easier to make the best use of 'learner output' and assess the learners appropriately. The developed rating scale is a sample and is created for assessing fifth grade students, but some changes can be made to make it fit an appropriate level and this will be difficult for teachers to do if adequate training is not conducted. A testing specialist has to conduct the training to help teachers get the point of the importance and the usefulness of using the rating scale and also be able to be prepare and adapt the scales themselves later on. A great deal of attention was paid to the type of categories that were included in the rating scale. However, this is only one side of the coin, as McNamara (2000) points out: rating can be objective only if the rating criteria are clearly and explicitly described and the raters are "trained carefully to interpret them" (p. 37).

4.4.2 Consequences of using the rating scale

The rating scale is intended to create fairness and minimize subjectivity in assessing students' performance. However, teachers, if not provided with some kind of help, guidance and supervision, may continue their assessment practices viewing each category and guideline the way they want and the issue of 'fairness as justice' mentioned by Kunnan (1999) will lose its value. The reason fairness is of such a concern for language testers is that "tests are very powerful instruments which can determine the future of individuals and programs" (Shohamy, cited in Kunnan, 2000). Though, classroom tests are considered low-stakes tests and do not affect students' immediate futures, a certain amount of attention should be paid to how they are constructed and assessed as the results of those tests may become 'self fulfilling' for students and unfair judgment "may have a damaging effect on the student's future performance" (Clapham, 2000, p. 151).

Another crucial consequence which may result from the effective or ineffective use of the rating scale is the washback effect that the new way of assessment may have on students and the learning process. To be able to adjust their teaching materials and activities with the assessment criteria or vice versa teachers need to be well aware of how to develop those criteria. In this case, the assessment will most probably have a positive/beneficial washback effect, given the explanation of beneficial washback which refers to interrelatedness of course content and objectives and those of the test (Hughes, 1989).

It will be difficult for teachers to implement the rating scale immediately after they are introduced to what kind of assessment tool it is. The implementation process needs to have support and consideration from the government, otherwise the teachers alone are unwilling to create additional work for themselves. Only those having keen interest in their job will try to do this. Therefore, it is necessary for government and school administrations to come together to make the implementation of such an assessment instruments possible.

Chapter Five: Conclusion

As stated in the introduction, the aim of this research was to develop a rating scale for assessing speaking in the fifth grade of Armenian secondary schools and test its usefulness in classroom assessment. Though many factors were considered while developing the rating scale and the teachers who used it in their classrooms showed interest towards using it, it has a long way to go to be implemented in classrooms as a permanent assessment tool. The following sections will cover several issues which need to be addressed: (a) limitations of the study, (b) suggestion for further study and (c) contribution of the study.

5.1 Limitation of the study

The first and the most obvious limitation of the study is the limited number of participants; therefore, the findings cannot be generalized to all the schools of Yerevan and the whole country. It may not be applicable to all schools of Armenia because many schools have a problem in teaching English, let alone mastering the assessment of it. The assessment of speaking is particularly difficult, which is the reason that in many cases discrete-point tests are used as they are easy to administer and score. In the case of assessing oral performance, the main concern is that "it can be difficult to set up and control" (Nunan, 2004, p. 141). Armenian teachers, however, mainly base their grades on speaking relying on their experience and overall impression. The fact that they do not consider certain aspects of language makes them think that it is a rather easy task to do.

The second limitation is that the researcher provided the teachers the prepared rating scale and made the necessary changes to it before every lesson so the teachers were responsible only for using it. Therefore, the reason the teachers showed willingness to use the rating scale later and assured that it is applicable and doable, may be because they did not have to modify it for each lesson themselves. If they were asked to deal with the modification part as well, most probably they would have been able to do this and in so doing would have identified some problems themselves.

5.2 Suggestion for further study

This study aimed at testing the effectiveness of assessing students' oral performance using a rating scale with guidelines and descriptors for each score. However, it included only teachers' reports on the usefulness of this way of assessment. A future study may be conducted including students' feedback and their perceptions of assessment in the classroom generally and of this way of assessment. This could help triangulate data and give a clearer picture of the issue. Moreover, if both students and teachers are included, a more comprehensive study can be carried out observing the washback effect of assessment with a rating scale on teaching and learning processes.

Other studies may be conducted to explore the assessment of other skills and other levels as well. The adaptation of this scale for assessing other skills is also possible because it covers the basic aspects and several necessary changes can make it applicable for other contexts. This would be possible to do provided that teachers are trained and feel comfortable with designing and using rating scales for different purposes based on the curriculum and the objectives they have. Another important issue for teachers to consider for later use of the rating scale is that besides the curriculum and the objectives of the course they have to have a clear identification of the benchmarks of the scores to be assigned to the students at different levels.

5.3 Contribution of research

This study of speaking assessment in Armenian secondary schools is a contribution to the assessment done in Armenian secondary schools. As assessment is not developed very well in Armenia and the system used is not flexible enough to reveal students' real knowledge and abilities, this research may serve as the first step in making relevant changes.

This raises teachers' awareness of the methods and effectiveness of using criteria for assessing the students rather than basing their assessment on their overall impression, which can easily be considered subjective and unreliable. Last but not least, this may also attract the attention of the Ministry of Education regarding the way assessment is carried out in Armenian secondary schools. It is important to have the attention of the Ministry because it has the power of making decisions and relevant changes to improve education in Armenia. The fact that a properly designed assessment tool may lead to better education and consequently a better generation of language speakers may make them think about improving classroom assessment and take appropriate measures.

References:

Alderson, C.J. and Banerjee, J. 2002. Language testing and assessment (Part 2). Language Teaching, Volume 35, Issue 02. April 2002. pp79-113.

Bachman, L 1990. Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Bachman, L. 1991. What does language testing have to offer? TESOL Quarterly, 25(4), 671–704.

Bachman, L. & Palmer, A. 1996. Language Testing in Practice. Oxford: Oxford University Press.

Berry, R. 2003. Alternative Assessment and assessment for learning. Hong Kong institute of education. From http://www.aqa.org.uk/support/iaea/papers/berry.pdf

Black, P. and William, D. 1998. Inside the Black Box: Raising standards trough classroom assessment. London Department of Education and Professional Studies; King's College London.

Bostwick, R. M. & Gakuen, K 1995. Evaluating Young EFL Learners: Problems and Solutions. In Brown, J.D. and Yamashita, S. O. (eds.) JALT Allied Materials Language Testing in Japan Tokyo: The Japan Association for Language Teaching: 57-65.

Brindley, G. Classroom based assessment. Macquarie University (Australia).

Brindley, G. 2001. Assessment. In The Cambridge Guide to Teaching English to Speakers of Other languages. Ed. Ronald Carter and David Nunan. Cambridge: Cambridge University Press.

Carrasquillo, A. L. 1994. Teaching English as a second language: A resource guide. New York: Garland Publishing.

Cheng, L. 1997. How Does Washback Influence Teaching? Implications for Hong Kong. University of Hong Kong, ©1997 l. Cheng, Language Education Vol. 11, No. 1, 1997.

Clapham, C. 2000. Assessment and Testing. Annual Review of Applied Linguistics. Cambridge: Cambridge University Press. From http://journals.cambridge.org/bin/bladerunner?REQUNIQ=1113548585&REQSESS=5441711&118200REQEVENT=&REQINT1=168362

Cohen, A.D. 1994. Assessing Language Ability in the Classroom. Second Edition. New York: Heinle and Heinle.

Davidson, F., C. E. Turner and A. Huhta. 1997. Language testing standards. In C.

Clapham and D. Corson (eds.) *Language testing and assessment, Vol. 7.*

Douglas, D. 1994. Quantity and quality in speaking test performance. *Language Testing, 11*(2), 125–44.

Farhady, H. 1999. Ethical Aspects in Language Testing. Iran, University of Science and Technology. Modarres, Vol. 3, No. 2, Summer 1999.

Farhady, H. 1995. Research Methods in Applied Linguistics. Payame Noor University.

Genesee, F. & Upshur, A. 1996. Classroom-based Evaluation in Second Language Education. Cambridge: Cambridge University Press.

Harutyunyan, N. 2002. An overview on the assessment and examinations. In A study based on the teaching of English language in the secondary schools of Armenia 2001-2002 academic year AELTA, Armenia.

Hughes, A. 1989.Testing for Language Teachers. Cambridge: Cambridge University Press.

Ingram. D.E. & Wylie, E. 1993. Assessing Speaking Proficiency in the International English Language Testing System. In Douglas & Chapelle (Ed.) A New Decade of Language Testing Research.

Kunnan, A. J. 1999. Recent Developments in Language Testing. Annual Review of Applied Linguistics, 19. 235-253. 1998.

Kunnan, A. J. 2000. Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando. Cambridge: University of Cambridge Local Examination Syndicate and Cambridge University Press.

Lewkowicz, J. A. 1997. Investigating authenticity in language testing. Unpublished PhD dissertation, Lancaster University, Lancaster.

Lewkowicz, J.A. 2000. Authenticity in language testing: some outstanding questions. *Language Testing, 17*(1), 43–64.

Lewkowicz, J. & Nunan, D. 2004 Task-Based Assessment for Learning Hong Kong: Hong Kong Education and Manpower Bureau.

Lumley, T. & Qian, D. 2001. Is speaking performance assessment based mainly on grammar? Paper presented at the Language Testing Research Colloquium, St Louis.

Luoma, S. 2004. Assessing Speaking Cambridge: Cambridge University Press.

Meiron, B. & Schick, L. 2000. Ratings, raters and test performance: an exploratory study. In A. J.Kunnan (Ed.), *Fairness and validation in language assessment* (Studies in Language Testing Series, Volume 9, pp. 153–76). Cambridge: UCLES Cambridge University Press.

McNamara, T. 2000. Language Testing Oxford: Oxford University Press.

McNamara, T.2001. The challenge of speaking: research on the testing of speaking for the new TOEFL. The University of Melbourne, Shiken: JALT Testing & Evaluation SIG Newsletter Vol. 5 No. 1 July 2001 (p. 2 - 3), from http://www.jalt.org/test/mcn_1.htm

Muradyan, K. 1999. Relationship Between Reading in English and L1 and L2 (Armenian and/or Russian) Language Proficiencies. American University of Armenia, Yerevan.

Nakamura, Y. 1993. Measurement of Japanese college students' English Speaking ability in a classroom setting. International Christian University, Tokyo.

Nunan, D. 2003. Practical English Language Teaching. The McGrow – Hill Companies, Inc.

Nunan, D. 2004. Task-Based Language Teaching Cambridge: Cambridge University Press.

Pavlou, P. 1997. Do different speech interactions yield different kinds of language? In A. Huhta,V. Kohonen, L. Kurki- Suonio & S. Luoma (Eds.), *Current developments and alternatives* 110 *in language assessment.* (pp. 185–201) Jyväskyla: University of Jyväskyla.

Rawls, J. 1971. A theory of justice. Oxford: Oxford University Press.

Shohamy, E., Donitsa-Schmidt, S. & Ferman, I. 1993. Test impact revised: washback effect over time. Tel Aviv University.

Shohamy, E. 1993 The power of test: The impact of language testing on teaching and learning. *NFLC Occasional Papers*, June 1993.

Shumin, K. 2002. Factors to Consider: Developing Adult EFL Students' Speaking Abilities. In Richard and Renandya (eds.) *Methodology in Language Teaching An Anthology of Current Practice* Cambridge: Cambridge University Press.

Smith, L. P. & Ragan, T. J. 1999. Instructional Design. John Wiley & Sons Inc..

Sook, K. H. 2003. The Types of Speaking Assessment tasks used by Korean Junior Secondary School English Teachers From http://www.asian-efl-journal.com/dec_03_sub.gl.htm

Spolsky, B. 1985. Fourteen Years on – Later Thoughts on Overall Language Proficiency. In Hughes, A. and Porter (eds.), Current Developments in Language Testing. London: Academic Press.

Stiggins, R. 2002. Assessment crisis: the absence of assessment for learning. Phi Delta Kappan, 83 (10).

Taylor, L. Issues in Speaking Assessment Research. Cambridge: Cambridge University Press From http://www.cambridgeesol.org/rs_notes/0001/rs_notes1_5.cfm#top

Venema, J. 2002. Developing Classroom Specific Rating Tests: Clarifying teacher assessment of oral communicative competence. Nagoya Institute of Technology. From http://www.jalt.org/test/ven_1.htm

Weir, C. 1993. Understanding and Developing Language Tests. Prentice Hall.

Appendix A[1]

Assessment Tasks: Set D
Task 2 – Telling your friends about Singapore

[2]General Criteria for assessing speaking	Task specific criteria
Content – demonstrating ➢ relevance of ideas to the topic ➢ appropriateness of ideas ➢ substantive coverage ➢ creativity and originality of ideas	**Content** ➢ Attraction, what is there, how long spent there, with whom and what you liked ➢ No irrelevant, inappropriate content ➢ Substantive content ➢ Provides additional information creativity
Organization – demonstrating Coherence through ➢ Using appropriate rhetorical patterns (e.g. narration, description, classification, comparison and contrast) ➢ Providing openings and closings as needed ➢ Presenting logically with appropriate examples/supporting details as needed ➢ Cohesion through effective use of appropriate repetition, connectives (e.g. conjunctions, time/order words), referencing, restatement, summarizing and tense consistency etc.	**Organization** Coherence ➢ Exploits the rhetorical pattern of description ➢ Clear opening and closing statements (e.g. I'm going to tell you about…) ➢ Points logically sequenced and combined ➢ Cohesion ➢ Coordination (e.g. I really liked it because…) ➢ Connectives (e.g. You can also go swimming)
Communicative Strategies – demonstrating ➢ Effective audience awareness (e.g. adjusting or modifying language/speech to suit audience needs, repeating and restating) ➢ Effective oral interaction strategies (e.g. agreeing, politely disagreeing, seeking clarification, clarifying, interrupting/interjecting, questioning, restating, summarizing, turn taking) ➢ Effective use of vocal features (e.g. eye contact, gesture, body movement and posture)	**Communicative Strategies** ➢ Effective use of voice and body language to make report interesting for the audience ➢ Uses examples to illustrate a point (e.g. …such as swimming and water sports; including sight-seeing)

[1] Adopted from: Lewkowicz, J. & Nunan, D. 2004 *Task-Based Assessment for Learning* Hong Kong: Hong Kong Education and Manpower Bureau

[2] This is a suggested list of general criteria for assessing speaking. Teachers might like to consider adapting it for use in their own classrooms.

Assessment Tasks: Set D
Task 2 – Telling your friends about Singapore

Genre and Task Requirements – demonstrating ➢ Adherence to the relevant requirements of different genres of speaking (e.g. story-telling, oral presentation. Public speaking, interview, conversation) ➢ Adherence to task requirements (e.g. authenticity, time limit)	**Genre and Task Requirements** Genre ➢ Speaks in a friendly and informal way Task ➢ Speaks for approximately 90 seconds ➢ Covers required content
Pronunciation and Fluency ➢ Clear and accurate pronunciation ➢ Audible articulation ➢ Smooth, confident delivery marked by ➢ Appropriate intonation ➢ Appropriate pauses and word stress ➢ Few hesitations ➢ Use of contracted forms ➢ Appropriate use of vocal features such as pitch, pace and tone	**Pronunciation and Fluency** ➢ Clear and accurate pronunciation ➢ Audible speech to maintain audience attention ➢ Appropriate intonation (e.g. with sentence tags - …okay?) ➢ Appropriate pauses and word stress ➢ Few hesitations ➢ Appropriate use of contracted forms (e.g. I'm going to tell you about) ➢ Effective use of vocal features (e.g. pitch, pace, tone)
Language and Style – demonstrating ➢ Appropriate range of vocabulary ➢ Effective choice of words ➢ Appropriateness of register for intended audience and purpose	**Language and Style** ➢ Vocabulary well-chosen and varied ➢ Language used appropriate for task and audience
Grammar – demonstrating ➢ Accuracy in grammar (e.g. subject-verb agreement, tense, modals, word order, prepositions, clause structure)	**Grammar** ➢ Effective use of past tense (e.g. we went there… we stayed, we had a gook time) ➢ Correct word order and subject verb agreement ➢ Pronouns (e.g. my friend and her mother) Connectives (e.g. I like it because…)
Visual Aids ➢ Appropriate use of visual aids for the intended purpose	**Visual Aids** ➢ (None needed or required, none assessed)

Appendix B
The developed rating scale

Criteria for Assessing Oral Performance of 5th-grade Students of Armenian Secondary Schools

Student: ______________________ **Date:** ______________________

1- poor (below minimal expected level), 2-needs improvement, 3-satisfactory, 4-good, 5-very good

General Criteria for assessing speaking skills	Task specific criteria
Grammar is accurate in using • tenses (different forms of present and past) • conditionals (first and second) • adjectives (the degrees of comparison) • nouns (plural forms, countable and countable) • passive voice • constructions with *used to* • reported speech	**Grammar** 1 2 3 4 5 Is able to • use and identify reported speech • switch from direct speech to indirect and vice versa freely • identify the difference between the past simple and past continuous • switch from one structure to the other appropriately
Vocabulary • uses previously learned vocabulary • uses words adequate to express intended meaning • avoids translations into the native language	**Vocabulary** 1 2 3 4 5 • uses appropriate vocabulary for the task and the topic • uses translations of the words in native language
Fluency the speech is • relatively fluent, without much hesitation • coherent and clear • easy to follow	**Fluency** 1 2 3 4 5 speaks • fluently, without much hesitation • clearly and coherently
Content • provides information relevant to the topic and the task • initiates topics/questions related to the task	**Content** 1 2 3 4 5 • the questions and the answers are relevant to the topic
Communicative Strategies • handles the topic of the conversation with confidence • is able to modify the speech to make himself/herself understood • uses body language adequately	**Communicative Strategies** 1 2 3 4 5 • has sufficient information about the topic and feels free to talk about it • is able to change or simplify the question to make it understandable • uses gestures and examples
Comprehension • Understands and is able to respond to the interlocutor's (peer's/teacher's) message	**Comprehension** 1 2 3 4 5 • understands the interlocutor's intended meaning and gives adequate answers

General comments:

Appendix C

The transcript of the recorded lessons (only fragments)

T: Ok, you are going to speak about your grandfathers and grandmothers.

S1: What about fathers? I don't write any father.

S2: My grandfather to be good at Georgian. He come from Georgia.

S2: "Mayrakaghak" (capital) how can we say.

T: Capital city.

S1: We lived there for four, four or five years. Then we came to Armenia and started to live here.

T: And?

S1: And then, then, er...

T: What's your favorite toy?

S1: My favorite toy now is Lego.

T.: Now speak about your family.

S2: Family, ok. We live here I, my mother and father. I have aunts and uncles. Some are in America and the other part is in Lebanon.

T: Have you ever been in America?

S2: No.

T: Are you going to visit your uncles in America?

S2: I don't know but, er..

T: Ah, they are visiting you and you don't need to go there.

S2: Yes!

T: Do you like your father?

S2: Sure...

T: Is your father very serious?

S2: My father? My father is not very serious 'artakinits' how can we say.

T: Appearance.

S2: Yes.

S3: Because he has mustache and er.

T: What are your favorite subjects?

S2: My favorite subjects are security survival English.

S3: I born in Russia,

T: I was born

S3: I very like Moscow.

Class: I very like (other students repeat and laugh to show that their friend made a mistake)

S3: I go there usually in summer I go to Rostov or Sochi. My childhood was very happy because every summer we go to our country and there are... I have there a lot of friends and my friends have horse and we ride on it and then I like eat pizza hamburger with sausage, cucumber Poland cheese and mayonnaise.

T: Ok what is your favorite book?

S3: My favorite book is ... how to say "koms" in English.

T: Duke

S3: "Duke Monte Cristo" which is 1200 pages.

T: Have you read it?

S3: Yes, the whole book.

T: What's your opinion about duke Monte Cristo? Do you think he is brave?

S3: In some ways.

T: Ok, what about his sweetheart.

S3: Oh, he is very unlucky because her sweetheart Mercedes are married with Ferdinand. They have, er, Ferdinand and Mercedes have a boy and they call them Albert. and Duke Monte Cristo and Albert and Ferdinand meet in Rome and then....

T: What is your favorite cartoon?

S3: My favorite cartoon is ... in Olymp

T: Why do you like it.

S3: I like the heroes because they are very clever and they are easy can foolish all people.

S4: I born in Yerevan.

T: I was born ...

S4: I was born in Yerevan in 1993. I used to play with those house. My favorite food is pizza, only pizza, I like

T: Who is your favorite hero?

S4: I like every one and I go to school, in school my favorite subject is mathematics, Armenian, English Russian.

T: Have you ever been to Russia?

S4: Yes, once.

T: Speak about that.

S4: I don't remember because I was very small.

S5: I was born in country, then I four five years live in Russian and then come in Armenian.

T: You came to Armenia then? Speak about your family.

S5: My family is I like my father

T: Do you like your sister?

S5: No!

T: Why? What do you think of her? Is she kind?

S5: No

T: Is she elder or younger?

S5: She is elder.

T: Does she study here.

S5: Yes.

T: Ok, you just don't like her and you avoid speaking about her.

S6: I was born in 1994, I live in Yerevan, I was born in Yerevan and I went to Moscow for three years. There was very happy and I went Yerevan I go to school and ...

T: Do you watch films? Who is your favorite actor?

S6: I like Chess Lee

S7: What about Bruce Lee.

T: What about your favorite singer?

(A long pause)

T: Don't you have a favorite singer?

S6: No.

Appendix D
The transcript of interview 1

Teacher A

1. **Do you think this is a good tool of assessment for everyday use?**

At first when I saw this sheet I thought that it was difficult to use and that is it was time consuming but when I started using it, I thought of ways how to use it so that both my students and me could benefit it. Therefore, I decided to write all the categories included in the rating sheet on the board and explain the students what each of them means and how they will be addressed while assessing them. For example, for grammar I told them that they should use the right tense form while speaking and for the vocabulary; I told them that they have to use as much of the newly learned vocabulary as they can; and the rest of the categories accordingly. They got very interested in it and became very motivated. They got involved in this process so much that they even started participating in the assessing of their peers. This raises the students' awareness of not only assessment process but also they start to be more careful with their speech to be able to achieve the highest score in each category. The focus was no more on accuracy or fluency other factors were also considered. Moreover, when I assigned them the average score of all five categories they knew why they got that certain score.

2. What difficulties did you encounter while using the rating sheet?

I had difficulties with communicative strategies as the students are used to stand and look at the walls while they tell something. Therefore, it was strange for students to be asked to keep eye contact, etc while speaking. But, I taught them several strategies and little by little they got the point and succeeded.

3. So you did not only use the sheet for assessing but also for teaching something.

I did teach them those communicative strategies because I think that it is very useful for them and it will be easier for them to communicate later on. Of course, it is very difficult for them to keep eye contact because they have to look somewhere when try to concentrate, but they try hard because they know that it will also contribute to their final grade.

4. What do you think about the scores provided in each category? Did you have any difficulties identifying where the student stands between the scores 1 – 5?

At first I thought that .it would be better if there were more scores but then I thought that it was fine to have 1 – 5 because, both me as a teacher and the students are used to that range and it is better to have it that way.

5. Do you think using this rating sheet was more/less time consuming than using the assessment method you are used to?

No, no. I wrote it on the board during the break and so I did not spend class time on it. Later, which surprised me very much; my students themselves wrote the categories on the board before the lesson. They got really involved in the process and showed more interest both towards learning and the way they were assessed. So the matter here is just getting used to it.

After getting the way how to use it, it is not time consuming. I find it very useful and I am going to use it later as well.

6. Do you think your students found the information provided useful? In what ways?

The students first had negative reaction to communicative strategies, as it was not easy for them to start using those strategies at once. But when they got used to it they found it very useful. They were really happy that they had the opportunity to get involved in the assessment process and they knew what their score meant.

7. Was there any difference in the scores students got from the previous way of assessment and this assessment method?

There was a little change because more aspects of spoken language are considered the students scored lower than before. I assessed only their overall performance before. But the difference is not significant.

8. Was there a case when the students got lower grades before and have higher grades when graded by this new assessment sheet?

I think there was a little bit change towards lower grades as more details of performance are considered here, before this, however, I just evaluated their overall performance. There was a case when a girl, who was very reserved and was not active during the lessons, was motivated by it and volunteered to participate and show that she also can. This helped her to be involved in the lesson and as she liked this process and this way of assessment, she came to the class prepared so that she may perform accordingly to get higher scores.

9. What do you think are the advantages of using this means of assessment?

It is very useful it raises the consciousness of the teacher and the students and using this sheet students will be able to communicate better. This sheet helps the teacher answer the students' questions/students' "why"s after having their grade. Usually the students ask "why 4" "why 3" now they know that they deserve certain grade because of several reasons. While student(s) speak both me and the rest of the students note down the mistakes that the person makes without interrupting the students and then we have the chance to both discuss the mistakes and also give reason for a certain grade.

10. So can I assume that this is not only assessment process but also learning process?

Certainly.

11. Do you think you will use such an assessment sheet in future? Why? Why not?

Yes, and as I mentioned I would like to keep a copy of the sheet to use it later. My students have the categories in their exercise books and they like the process so I want to use it. As it really helps both teachers and students. Only one should find a way to use it effectively.

12. What changes would you suggest to make in the rating sheet to make it more practical for everyday use?

I haven't thought about it before. It would be better to have more range in the scores but given the fact that the usual assessment method is kept, it is ok.

13. You said there was a problem with communicative strategies. Do you think it is better to take that category out of the sheet not to create problems?

No, no, no. I think it is a step forward. It is better to keep it and try to teach it students because it is very important in communicating in a foreign language. As communication is essential for learning a language, communicative strategies are also essential. I do not think it is a good idea to take this category out, it is better to teach students those strategies.

Appendix E
The transcript of interview 2

Teacher B

1. **Do you think this is a good tool of assessment for everyday use?**

I think it is a good one. The only matter is that it is a bit time consuming. It took me three to five minutes to assess each student every day. However, I think that if one gets used to it, it will be easier to use and consequently it will be more effective.

2. What difficulties did you encounter while using the rating sheet?

As I said, it is time consuming and as it was new for me, I had to get accustomed to using it. Little by little while using it I spent less time on it. Therefore, if it becomes a part of everyday assessment permanently the issue of the rating sheet being time consuming will not be a problem any more.

3. What do you think using this rating sheet was more/less time consuming than using the assessment method you are used to?

Generally, I rely on my overall impression while assessing students so it is less time consuming than the new rating sheet. So again, the only problem is the time, which can be overcome over time.

4. Do you think you students found the information provided useful? In what ways?

They were excited to know that they were graded in a different way. They asked, "what is this" and I explained what categories were included there. They got motivated to perform accordingly to get higher grades.

5. Was there any difference in the scores students got from the previous way of assessment and this assessment method?

I wouldn't say there was a big change in the grades students got when I used the usual assessment method and the grades that they got from this rating sheet. There was a little difference and the students got lower marks. While using this sheet I paid much attention on how they use the vocabulary of that day which I would not consider so much before thinking that they will learn it later.

6. What do you think are the advantages of using this means of assessment?

The advantage of using this sheet is that the teacher becomes more skilled to assess the students and becomes clearer in his/her assessment paying attention to different aspects. Though before having this assessment sheet I used holistic approach, I find this useful as in this case students are more conscious of the way they are assessed which means that they are willing to study better. It fosters students learning as well because if they know that they are going to be assessed according to their knowledge of grammar, vocabulary and other skills they prepare more carefully.

7. Do you think you will use such an assessment sheet in future? Why? Why not?

Yes, it was an effective way of assessment, in my opinion and I think I will use such a sheet in future. Later, if I have a choice, I will prefer to use this rating sheet though I like holistic assessment more than analytic. However, as this way of assessment reveals more students performance I prefer it.

8. What changes would you suggest to make in the rating sheet to make it more practical for everyday use?

I am not sure but I think it would be better to reduce the number of criteria to make it shorter and less time consuming. I think that though grammar is very important aspect of the language, at this level and especially for oral performance it could be paid less attention, as at this level there is the issue of getting the students' fluency in oral communication. At this point acquiring fluency is more important than accuracy as they know the grammar just while speaking they misuse it and their grades suffer from it.