

EVALUATION OF THE RELIABILITY OF
TWO GRADING SYSTEMS FOR WRITING ASSESSMENT AT
ANADOLU UNIVERSITY PREPARATORY SCHOOL

A THESIS PRESENTED BY
NESRIN ORUC
TO THE INSTITUTE OF ECONOMICS AND SOCIAL SCIENCES
IN PARTIAL FULLFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTERS OF ARTS
IN TEACHING ENGLISH AS A FOREIGN LANGUAGE

BILKENT UNIVERSITY

JUNE 1999

T 89659

ABSTRACT

Title: Evaluation of The Reliability of Two Grading Systems for Writing Assessment at Anadolu University Preparatory School.

Author: Nesrin Oruc

Thesis Chairperson: Dr. Patricia N. Sullivan
Bilkent University, MA TEFL Program

Committee members: David Palfreyman
Michele Rajotte
Bilkent University, MA TEFL Program

Evaluating student writing is always a difficult task for English teachers. Not only it is time consuming to read a seemingly endless stack of compositions, but it is also difficult to write a grade in red ink on the work of a student which represents so much of her/himself, and to put our students into categories such as “A” student, “C” student, or “F” student brings up the issue of the most effective ways of assessing student writing.

Assessing students’ writing has always been accompanied by arguments between writing instructors at Anadolu University. This is not only because of the long hours spent on grading compositions, but also because of the inconsistencies among the grades and the graders. The main reason for this problem is that at Anadolu University a Standard Grading Format (SGF) has not been used for writing assessment. Instead writing instructors have used their individual grading system to evaluate papers.

The main purpose of this research study was to examine writing instructors’ individual approaches to assessing writing and then to determine whether the use of a holistic scoring scale would result in an increase in the reliability of the writing assessment at Anadolu University Preparatory School.

The participants were six writing instructors from Anadolu University Preparatory School teaching writing to different levels. Half of the instructors were male, half were female. The instructors had between 1 and 10 years experience in language teaching and they were all non-native speakers. The process was begun by giving 15 papers chosen to represent different grade levels from the Anadolu University Preparatory School placement exam to instructors to grade in their accustomed manner. Then, after a training given an SGF using a five band holistic scoring scale which was adapted to Anadolu University Preparatory School, instructors were again given the same 15 papers and were asked to use the new SGF to grade the papers again.

Data were analyzed first by a One-way ANOVA in order to find the mean scores of each system. Then the Student T-Test was used to compare the mean scores of the systems. The results of the One-way ANOVA indicate that there is significant relationship between the grades given to the same paper by five different instructors before and after the training which means both of the systems were reliable within themselves. On the other hand, the Student T-Test results indicate that there is a large difference between the scores given to the same papers by the same instructors with two different writing assessment systems. Consequently, no clear judgement can be made as to which system is superior.

With regards to the grades given by the instructors for individual papers, the results of qualitative analysis indicate that inconsistencies arise from individual instructors' writing assessment practices and that this may be lessened with holistic scoring.

BILKENT UNIVERSITY
INSTITUTE OF ECONOMICS AND SOCIAL SCIENCES
MA THESIS EXAMINATION RESULT FORM
JULY 31, 1999

The examining committee appointed by the Institute of Economics and Social
Sciences for the thesis examination of the MA TEFL student

Nesrin Oruc

has read the thesis of the student.

The committee has decided that the thesis of the student is satisfactory.

Thesis Title: Evaluation of the Reliability of Two Grading Systems
for Writing Assessment at Anadolu University
Preparatory School.

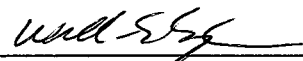
Thesis Advisor: Dr. William E. Snyder
Bilkent University, MA TEFL Program

Committee Members: Dr. Patricia N. Sullivan
Bilkent University, MA TEFL Program

David Palfreyman
Bilkent University, MA TEFL Program

Michele Rajotte
Bilkent University, MA TEFL Program

We certify that we have read this thesis and that in our combined opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Masters of Arts.



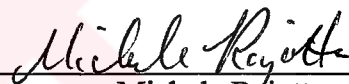
Dr. William E. Snyder
(Advisor)



Dr. Patricia N. Sullivan
(Committee Member)

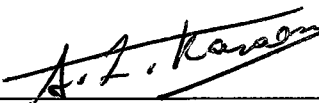


David Palfreyman
(Committee Member)



Michele Rajotte
(Committee Member)

Approved for the
Institute of Economics and Social Sciences



Ali Karaosmanoglu
Director
Institute of Economics and Social Sciences

ACKNOWLEDGEMENTS

I would like to express my special thanks to my thesis advisor, Dr. William E. Snyder, for his helpful guidance, and continuous encouragement throughout my research study.

I am thankful to Dr. Patricia N. Sullivan, Dr. Necmi Aksit, David Palfreyman, and Michele Rajotte, who provided me with invaluable feedback and recommendations.

I am deeply grateful to the Director of the Preparatory School of Anadolu University, Prof. Dr. Gül Durmuşoğlu Köse, who provided me with the opportunity to study at MA TEFL at Bilkent University.

I owe the deepest gratitude to the instructors, Aynur Baysal, Ayşegül Aktaş, Handan Girginer, Bülent Alan, Erol Kılınç, and Naci Afacan, who participated in the study. Without them, this thesis would never have been possible.

I am sincerely grateful to all my MA TEFL friends, especially Ayfer Sen for being so cooperative throughout the program.

Finally, I must express my deep appreciation to my dear family, who have always been with me and supported me throughout.



To my family for their never ending support and love

and

to the stars for brightening the nights of long study hours...

TABLE OF CONTENTS

LIST OF TABLES.....	x
LIST OF FIGURES.....	xii
CHAPTER 1 INTRODUCTION.....	1
Background of the Study.....	5
Statement of the Problem.....	6
Purpose of the Study.....	7
Significance of the Study.....	7
Research Questions.....	8
CHAPTER 2 REVIEW OF THE LITERATURE.....	9
Purposes of Assessment.....	10
Methods Used for Writing Assessment.....	13
Analytic Scoring.....	15
Holistic Scoring.....	18
Comparing Analytic and Holistic Methods.....	21
Problems of Writing Assessment.....	22
Fairness.....	22
Training.....	26
CHAPTER 3 METHODOLOGY.....	29
Introduction.....	29
Participants.....	29
Materials.....	30
Procedures.....	31
Data Analysis.....	33
CHAPTER 4 DATA ANALYSIS.....	34
Overview of the Study.....	34
Participants.....	34
Materials.....	35
Data Analysis Procedures.....	36
Results.....	37
Grades Given To The Papers After Placement Exam.....	38
Comparison Of The Means Of The Grades Given To Each Papers.....	39
Instructors' Grades Given To 15 Papers.....	42
The Grades Assigned to Paper Number 4 By All Instructors.....	43
The Grades Assigned to Paper Number 7 By All Instructors.....	43
The Range Of The Grades.....	44

The Grades Given To The Same Papers By The Same Instructors After The Training On The Holistic Scoring System.....	47
One-way ANOVA Variance Analysis For The Holistic Scoring System.....	48
One-way ANOVA Variance Analysis For The Old Analytic System.....	49
Paired Samples Statistics.....	50
Paired Samples Correlations.....	50
Paired Samples Test.....	51
Scores Given To The Same Papers Before And After The Training.....	52
CHAPTER 5 DISCUSSION OF RESULTS AND CONCLUSIONS.....	54
Overview of the Study.....	54
Summary of Results and Conclusions.....	54
The Reliability of the Old Analytic Writing Assessment System Used at Anadolu University Preparatory School.....	55
The Reliability of the New Holistic Scoring System Used at Anadolu University Preparatory School.....	56
The Difference Between the Grades Given to the Papers in the Two Different Scoring System.....	57
Limitations of the Study.....	58
Curricular Implications.....	59
Scoring Criteria.....	59
Effects On Instructors.....	60
Training.....	61
REFERENCES.....	63
APPENDICES.....	66
Appendix A: Student Papers.....	66
Appendix B: Holistic Scoring Scale.....	86
Appendix C: Instructors' Grades.....	88

LIST OF TABLES

<u>TABLE</u>		<u>PAGE</u>
1	Grades Given to the Papers After Placement Exam.....	38
2	Comparision of the Means of the Grades Given to Each Paper..	39
3	Instructors' Grades Given to 15 Papers.....	42
4	The Grades Assigned to Paper Number 4 By All Instructors.....	43
5	The Grades Assigned to Paper Number 7 By All Instructors.....	43
6	The Range of the Grades.....	44
7	The Grades Given to the Same Papers By the Same Instructors After the Training on the Holistic Scoring System.....	47
8	One-way ANOVA Variance Analysis for the Holistic Scoring System.....	48
9	One-way ANOVA Variance Analysis for the Old Analytic System.....	49
10	Paired Samples Statistics.....	50
11	Paired Samples Correlations.....	50
12	Paired Samples Test.....	51
13	Scores Given to the Same Papers Before and After the Training.....	52

LIST OF FIGURES

<u>FIGURE</u>		<u>PAGE</u>
1	An Example of an Analytic Scale.....	16
2	An Example of a Holistic Scale.....	19



CHAPTER 1

Introduction

Evaluating student writing is a difficult task for English teachers. Not only it is time consuming to read that seemingly endless stack of compositions, but it is always difficult to write a grade in red ink over the student's work, which represents so much of him/herself.

As teachers we ask all the students in our classes to do certain assignments, to write certain kinds of essays, and we assess them. After assessing the students' papers we put our students into categories such as "A" student, "C" student, "F" student. Putting students into these categories raises the issue of effective assessment for institutional and instructional purposes.

Assessment of writing ability, however is an important task, not only for teachers but also for students. The grades found after the assessment are the record of a teacher's evaluation of each student's work. It is important for students that teachers should know more about assessing writing because the decisions they make about grades affect the students' lives and education (Williams, 1996; Brown, 1996; White, 1989). Teachers also should understand that different writing tasks require different kinds of assessments and qualitative forms of assessment can be sometimes more powerful and meaningful for some purposes than quantitative measures. Another point is that the assessment should be seen as a means to help students learn better, not a way of unfairly comparing students (Miller, 1997).

Writing assessment can serve to inform both the individual and the public about the achievements of students and the effectiveness of teaching. On the other hand, if the test has been designed poorly and has been assessed poorly, then this

can be enormously harmful to our students as human beings and as learners.

Students can be demotivated (Brown, 1996).

To be better teachers, teachers should clearly understand what is involved in evaluating writing. Three factors in assessing writing should be important for writing teachers. The first is the validity of the assessment. Validity is related to matching what teachers are measuring to what they are teaching and to what their assignments ask students to do. The second is reliability which describes the degree of consistency from one evaluation to another (Jacobs et.al., 1981; Hughes, 1989). Last is the amount of time spent on grading papers (Dinan & Schiller, 1995; Williams, 1996; Grabe & Kaplan, 1996).

Assessment is important for a range of educational stakeholders, including teachers, students, parents, employers, governments, and taxpayers (Williams, 1996; Lloyd-Jones, 1987; Dinan & Schiller, 1995; Brumfit, 1993). According to Brumfit (1993), assessment is important for learners, because they want to know how they are progressing, and they want some formal feedback, as well as informal comments and encouragement from their teachers. The importance for parents and employers is that they want to know if the learners are receiving worthwhile instruction. Finally, the importance of assessment for governments and taxpayers is that they want to know that teachers are not wasting precious resources through self-indulgence or laziness.

When conducted properly, assessment can have a positive impact on teaching, learning, curricular design, and student attitudes (Brown, 1996). Although this may seem very clear and achievable, it is generally one of the most difficult tasks of institutions. Yancey (1997) divides the rights and responsibilities of

conducting “sensitive and purposeful” writing assessment into four categories: students, faculty, administrators, and legislators.

According to her, teachers’ responsibility is to play key roles in the design of writing assessments, including creating writing tasks and scoring guides (Yancey, 1997). In order for instructors to meet these responsibilities in writing assessment, they should receive instruction that helps them use clear and appropriate criteria to assess a piece of student work. In short , the training of assessors is essential (Yancey, 1997).

At the same time, trained instructors should apply all of the standards that have been set for the particular testing program. For that reason, instructors should practice scoring a range of student writers’ papers throughout the scoring process. Although training teachers is an important task in writing assessment, the methods that a trained teacher will use is as important as the training. Unfortunately, there is no formula for the most successful method of assessment; each institution needs to define its philosophy for assessment clearly. In other words, many different types of assessment can work as long as the institution and/or the instructors understand why they assess in the way they do and how they believe these assessments will help students in the future (White, 1989; Yancey, 1997; Brown, 1996).

As times have changed, the methods that are used to assess writing papers have also changed. Different institutions are using different kinds of writing assessment systems according to their objectives and their students’ needs.

A growing number of teachers have moved away from the traditional method of evaluating student writing, which entails taking papers home, reading them, and

assigning a grade. Some teachers have become frustrated with this model not only because of the amount of time it takes, but also because of the message it sends to the students that one teacher is the audience for their papers (Miller, 1997 p.2).

When we analyze the traditional assessment practices, we can see that subjective scoring is used on short, open-ended questions or on open-ended tasks. Purely subjective scoring is to be avoided because the experiences of the instructor can influence the grade given to the student work. Since instructors have different experiences, the same student work may receive different scores from different instructors working at the same institution (Jacobs et.al. 1981).

White (1989) examines an interesting grade appeal case which he was involved in. Two teachers in different courses, one from a different university, graded a student's term-paper as "A" whereas the student received an "F" from her own teacher. White criticizes this situation with the following analysis " There was a certain grim comedy involved in this case of the much-used all purpose term-paper".

White also suggests that students in general believe that their writing papers are graded arbitrarily and inconsistently. This is because of the lack of clarity of grading standards used to grade the writing papers.

Background of the study

Anadolu University Preparatory School was established in 1998 and serves 650 preparatory students this year. The number of the instructors working at the school is 43 and seven of the instructors are native speakers of English.

Students are required to take a placement exam at the beginning of the term and are placed in levels of Beginner, Elementary, Lower-Intermediate, Intermediate, Upper-Intermediate and Advanced. The number of the students in each class varies between 20 and 25.

Anadolu University Preparatory School uses a skills-based syllabus, so each skill is taught separately. In each level students are given 4 hours of writing courses. Each instructor is responsible for one or two level's writing courses and is supposed to read the exams of his/her classes' writing alone. Only one instructor evaluates the papers.

The same problem pointed out in White's study is something that could happen easily at Anadolu University Preparatory School because a standard grading format is not used for writing assessment. One of the results of this situation is that different instructors may grade the same paper differently or the same instructor may give papers of same quality different grades.

Being a writing instructor is considered a kind of punishment given by the institution at Anadolu University Preparatory School. When an instructor is assigned to give writing courses, it means that the instructor will have to check drafts of each assignment, write comments on them, prepare exams and grade the papers. The biggest problem starts at that point because grading papers is difficult and the instructors feel uncomfortable about the inconsistencies that arise.

After each writing exam, instructors argue while trying to agree upon a grading format. The resulting format is generally not detailed. A general format covering 5 issues, Title, Punctuation, Spelling, Grammar, and Organization, has been developed by all writing instructors who teach writing at different levels. The same

format is used for the writing assessment of all levels, it is interpreted differently. For example, grammar is given 20 points, but the criteria from 0 to 20 are not specified, so writing instructors have no guidance in giving a grade in each category. Writing instructors at Anadolu University Preparatory School use a kind of internal grading system in order to evaluate papers, without asking if it is fair or not.

The grading format that is currently used has a number of other problems. The grading format that is developed after each exam is used for the assessment of each level despite the big differences between the levels of the students. The same format is sometimes used for the assessment of students' writing assignments too. Even though the aims of an assignment and an exam are different, writing instructors at Anadolu University Preparatory School use the same format for the assessment of two different purposes.

Statement of the Problem

Absolute objectivity is difficult in writing assessment. I believe a writing instructor should be objective about certain factors. But if the papers are graded by an internal grading system, the objectivity of the writing program can be questioned.

As was stated in the previous section one problem is that the assessment of the writing papers of a group is done only by the instructor who teaches writing to that group. A second problem occurs because a standard grading format is not used. In order to be objective the writing assessment of students' papers should be done not only by more than one instructor, but also with the use of a standard grading format.

Purpose of the Study

The purpose of my study is to examine both the writing instructors' standards in assessing writing and the inconsistencies that arise in the grading process.

I plan to develop a standard grading format for a writing assessment at Anadolu University Preparatory School and to assess whether the use of this standard grading format will result in an increase in the reliability of the writing assessment at Anadolu University Preparatory School.

Significance of the Study

Anadolu University is not the only institution where writing assessment is done without a standard grading format. All institutions, especially those where the writing instructors do not use a standard grading format for writing assessment, will benefit from this study.

As the objectivity of writing assessment is important for an institution, this study has a particular value for Anadolu University. The study will help not only instructors, but also administrators at Anadolu University to make a decision about what kind of assessment they want in order to assess their students' writing ability both objectively and reliably.

With the help of the writing instructors at Anadolu University a more detailed and useful format for grading writing papers can be developed at Anadolu University and not only instructors, but also students will benefit.

Research Questions

My study will address the following research questions regarding writing assessment at Anadolu University:

1. What is the reliability of the old analytic writing assessment system used at Anadolu University Preparatory School?
2. What is the reliability of the new holistic scoring system used at Anadolu University Preparatory School?
3. Is there a significant difference between the grades given to the papers in the two different scoring systems?



CHAPTER 2: REVIEW OF THE LITERATURE

Assessment has always been an important part of the overall teaching and learning process. Dalton (1993) defines assessment as an ongoing process that can take many forms, including standardized tests that can extend over time. These aim to document the quality of student work or the quality of an educational program. Assessment should reflect what we want students to know and be able to demonstrate over an extended period.

This simply serves to underline the fact that assessment, including assessment of written literacy, is a part of every teaching program. Scores found after the assessment are used to justify programs or change them, and scores may be used to penalize or reward students, so they have important public effects (Lloyd-Jones, 1987; Brown, 1996). The words “to penalize” and “to reward” are important since they give two different perspectives in the purpose of writing assessment.

For many of us, thinking about student writing includes grading. As highly trained professionals, we have literally spent years learning how to grade and be graded. For many of us, grades have meant whether we would get into a specific field, could choose a particular profession, or, in some cases, how we defined ourselves. Grades also have been the marks of our success. As we move into another role, however, we carry our personal and institutional experiences of grading with us (Miller, 1997).

Writing assessment, like the assessment of any other skills, has important consequences for students as well as for instructors (Grabe & Kaplan, 1996).

Assessing students' writing can influence students' attitudes to writing and their

motivation for future learning. Students can be demotivated by invalid and unreliable exam results and can be frustrated in their writing progress. Alternatively, students can be positively motivated by valid and reliable assessments of their writing. Their personal creativity may increase (Grabe & Kaplan, 1996; Brown, 1996; Dalton, 1993).

This chapter reviews the literature on writing assessment in three areas: purposes of assessment, methods used for writing assessment, and thirdly, problems of writing assessment.

Purposes of Assessment

Moursund (1995,1997) defines the three common purposes of assessment in education. According to Moursund, the first purpose of the assessment is to obtain information needed to make decisions. This information might be used by a variety of different stakeholders, such as students, teachers, parents, policy makers, and resource providers (See also Brown, 1995). These stakeholders often have different information needs and make differing types of decisions based on the assessment information received. An assessment designed to fit the needs of students, for example, may be different than assessment designed to meet the needs of teachers or policy makers.

Fulcher (1996) also defines the purpose of assessment to be gathering data for decisions and states the following:

In testing, we wish to gather evidence for a decision of some kind: to admit to an educational program, promote a student, identify as in need of remedial tuition, or to award a certificate. The evidence required may also differ given

the nature of the purpose and the seriousness of the decisions based on test results; that is, whether the assessment is “high stakes” or not” (p.92).

In another study Brumfit (1993) defines the purposes of assessment for decision making. According to Brumfit, assessment is made in order to learn whether the student has learnt all that was required by the syllabus, whether the teacher has taught the syllabus, and whether the student should go on to further study in the subject.

Moursund’s second purpose of assessment in education is to motivate the people or organization being assessed. Many other researchers have dealt with the subject (Brown, 1996; Grabe & Kaplan, 1996; Bailey, 1998) and decided that one of the important purposes of assessment is to motivate the student, teachers, and the administrators. In education, for example, it is often said that assessment drives the curriculum. Successful performances act as an affirmation to the students, teachers, school administrators, and other stakeholders. This motivates teachers to “teach to the test” and students to orient their academic work specifically toward achieving well on tests.

“Teaching to the test” is also known as “washback”, a topic which is covered in Bailey (1998). She describes “washback” as follows: “The effect a test has on teaching and learning” (p.3). Washback can be either positive or negative. For example, if a test requires students to spell a number of unusual words and their definitions, then the students will have to memorize the spelling and the meaning of these words that they will not use in the target language for their daily life communication needs. This is be an example of negative washback, because in a course promoting communication skills, memorizing the meanings and the spelling

of unusual words will not be useful for the students. Hughes (1989) also talks about washback and suggests that in order to achieve beneficial washback, direct testing, testing directly the skill that the instructor wants to test, should be used.

The third purpose of assessment in education is to emphasize the accountability of students, teachers, school administrators, and the overall educational system. For example, a school district's educational system might be rated on how well its students do on college entrance tests. Poor student performance may lead to major changes of administration in the school district (Moursund 1995,1997).

Jacobs, Zinkgraf, Warmuth, Hartfiel and Hughey (1981) identify the institutional purposes of a composition test as making assessments for decisions about entry to a school program and to gather test data for research into the nature of language. Accordingly, any single testing situation may serve multiple purposes, such as admissions and/or proficiency testing, providing information used for prediction, for placement, and for diagnosis.

Equally important are the classroom teachers' purposes. Classroom teachers may test composition for similar reasons, but more frequently they want to diagnose learners' needs, measure growth at the end of an instructional sequence, provide feedback to focus the learning efforts of student-writers, and to evaluate the efficacy of certain teaching methods or techniques (Jacobs et al., 1981).

Grabe and Kaplan (1996) also divide writing assessment into two contexts: the classroom context and the standardized testing context. The classroom context involves achievement assessment while the standardized context involves proficiency assessment. Classroom assessment can be used for diagnostic purposes.

In contrast, standardized assessment is primarily used to make proficiency judgements.

Brown (1996) states that in order to test appropriately, each teacher must be very clear about his/her purpose for making a given decision and then match the correct type of test to that purpose.

Writing instructors should have identified their purposes while preparing a writing test and assess their students according to these purposes. Another important point to be considered is that students should also be informed about these purposes and should be prepared according to them (Brown, 1996; Grabe & Kaplan, 1996; Bailey, 1998).

Methods Used for Writing Assessment

The way we assess our students' writing skills has changed during the last two decades. Traditional assessment practices utilized multiple-choice, norm-referenced or criterion referenced standardized achievement tests. In many instances, the use of these tests resulted in the narrowing of curricular offerings and contributed to the practice of "teaching to the test." Many of the older forms of achievement tests assessed student recall of facts and measured low-level thinking skills (Yancey, 1997).

In recent years, however, newer assessment instruments have been developed that incorporate opportunities for students to analyze, generalize, synthesize and evaluate. The curriculum reflected in these newer tests is more general in nature and emphasizes substance and application rather than factual recall. Students are asked to define, extrapolate, and respond creatively to test questions. Many institutions have implemented assessment processes that require active student participation

and/or collaboration in problem design and problem solving, which lead to a more integrative way of assessing student ability (Dalton, 1993).

As was stated above, according to the purpose of the course, the purpose of the teacher and the objectives of the institution, different kinds of writing assessment may work in different situations. Because of this, different people have studied different kinds of writing assessment methods suitable for their purposes (Dalton, 1993; Grabe & Kaplan, 1996; Brown, 1996).

Dalton (1993) divides writing assessment into four categories: alternative, performance, authentic, and portfolio assessment. Alternative assessment is any process, procedure, or product that will be used to assess or evaluate a student's knowledge of that particular subject matter area. Performance assessment provides one sample of a student's work under controlled conditions, such as making students write a story, essay, or article in a given period of time. In authentic assessment students are provided a prompt or a question and asked to write a response, description or argument within a limited time. The fourth and the last kind of assessment, which is portfolio includes samples of student written work.

Grabe and Kaplan (1996) introduce three different kinds of writing assessment: indirect writing assessment, direct writing assessment, and portfolios. Indirect writing assessment tests students' grammar, vocabulary, and written expression knowledge. Direct writing assessment tests students by having them produce an authentic piece of writing. Portfolio assessment, which is newer than the other two, is the opportunity for the students to bring writing samples that they have written outside the classroom for assessment.

Alderson, Clapham and Wall (1995) divide types of scales into two. In holistic scoring scales instructors are told to give a grade on the student's performance as a whole according to the scale that they use. In analytic scoring, instructors decide on the grade that they give according to several components of the analytic scoring scale that they use.

The major focus of this study is to compare the reliability of two different writing assessment systems. The first one is an analytic scoring system, which was being used until the time of the study and the second one is a holistic scoring system, which was suggested as a new scoring system to accompany this study. I will review these two writing assessment methods further below.

Analytic Scoring

Analytic scoring systems are divided into three types. The first, known commonly as the "point-off method" (Madsen 1983, p.120), focuses on writers' errors. The reader begins with an "A" grade or 100 percent and points are deducted for each error (Madsen, 1983; Dinan & Schiller, 1995; Hughes, 1989; Bailey, 1998).

The second type uses a rubric based on categories determined to be important for evaluating the quality of a piece of writing. Bailey (1998) describes the second type of analytic scoring as assessing students' performances on a variety of categories.

Dinan & Schiller (1995) discuss the process of designing and using a rubric as described below:

- Determining the specific features of writing that the institution wants to assess.
- Agreeing upon language to describe different quality levels of each selected feature, assigning a numerical score to each.

- Deciding how the selected features should be weighted.
- Using the rubric to read and assess the essays.
- Toting up the scores, taking “weighting” schemes into account, then deciding what to do with the results.

Madsen (1983) summarizes the second type of analytic approach as having the procedure described above. Points are given for acceptable work in each of several areas.

Mechanics	20%
Vocabulary choice	20%
Grammar & usage	30%
Organization	30%
TOTAL	100%

Figure 1: from Madsen, 1983 p.121.

On the other hand, other people give the numbers of the components differently. For example, Lloyd-Jones (1987) and Bailey (1998) give rubrics for analytic scoring systems consisting of five components. In Bailey’s list, it is believed that effective writing consists of five components, such as content, organization, language use, vocabulary and mechanics, whereas according to Lloyd-Jones (1987) the basis for some analytic scales are: ideas, organization, wording, flavor and mechanics.

These lists suggest the chief problem with analytic scales. To agree on the suitability of any scale and on the points assigned to the categories is difficult. Some

people do not want any points given for mechanics and others might consider mechanics essential.

A third kind of analytic scoring system for writing assessment is known as analytic-holistic scoring method. A report from the University of Texas writing center reviews this method with the following analysis:

Analytic holistic scoring can be used as means of informing both the scorer and students of general areas of high and low quality. Analytic holistic scoring follows the same procedures as focused holistic scoring, but the rubrics are more specific. The information provided by the rubrics in analytic holistic scoring is generally more useful to students, especially for beginners. Analytic holistic scoring rubrics can be used when only part of an entire paper is to be scored. Several analytic scores can be given to one paper. With a little practice a teacher will be able to read a paper one time and assign several analytic scores. Analytic rubrics are often used with short answer essays. Discussion of the analytic rubrics before and after the task can provide a vehicle of instruction” (University of Texas, 1997 p.4).

Analytic holistic scoring provides students and teachers with diagnostic information about students’ particular strengths and weaknesses and is desirable when students need feedback about their performance in key areas of their learning products (University of Texas, 1997).

Holistic Scoring

Another method that was developed to provide reliability is holistic scoring (Baker & Linn, 1992; Schwalm, 1993; Wolcott, 1996; Miller 1997; Hughes, 1997).

Schwalm (1993) describes this method with the following definition, “The word “holistic” means looking at the whole rather than at parts. Holistic scoring is a procedure for evaluating essays as complete units rather than as a collection of constituent elements” (p.9).

Parallel to analytic scoring, holistic scoring also has some steps which readers have to follow.

- Forming a group of teachers, probably several teachers teaching multiple sections of the same course.
- Creating criteria which might be a general criteria for all papers, or a set of criteria specific to each assignment.
- Norming to scale; to achieve a consensus among the group on how to apply the criteria to actual papers.
- Orchestrating the grading session, which means assigning one person to be administrator for the session (Miller, 1997).

An example of a holistic scoring system can be seen in Figure 2.

18-20	Excellent	Natural English with minimal errors and complete realization of the task set.
16-17	Very Good	More than a collection of simple sentences, with good vocabulary and structures. Some non-basic errors.
12-15	Good	Simple but accurate realization of the task set with sufficient naturalness of English and not many errors.
8-11	Pass	Reasonably correct but awkward and non-communicating OR fair and natural treatment of subject, with some serious errors.
5-7	Weak	Original vocabulary and grammar both inadequate to the subject.
0-4	Very Poor	Incoherent. Errors show lack of basic knowledge of English.

Figure 2: from UCLES International Examination in English as a Foreign Language General Handbook, 1997.

Three purposes for using holistic scoring system in writing assessment are enabling quick, reliable, and valid evaluation of student essays (Lloyd-Jones, 1987; Schwalm, 1993; Miller, 1997; Bailey, 1998).

Apart from the writers referred to, the advantage of a holistic scoring systems for reducing the amount of time needed to grade has also been reported by Hughes (1989). He claims that holistic scoring has the advantage of being very rapid. When experienced scorers are in the process of scoring they can judge a one-page piece of writing in a couple of minutes or even less. Moreover, it will be possible to score each piece of work more than once in a brief period of time. Bailey (1998) also

states that one of the advantages of a holistic scoring system is that it is a fast way of grading student papers. Miller (1997) in another study claims that holistic scoring system was originally developed in order to reduce the amount of time needed to grade.

The second advantage of a holistic scoring system is that it is a reliable way of assessing student essays (Schwalm, 1993; Miller, 1997; Bailey, 1998). Holistic scoring was originally designed to reduce the subjectivity of grading written work by relying on the initial, almost intuitive, reactions of a number of people in order to assign a grade. More importantly, holistic grading prevents the reader from concentrating only on grading; as a result, he or she reads student work more “naturally” as if she or he were reading any text (Miller, 1997).

Miller (1997) also states that many teachers are beginning to use the holistic grading model in order to help students better imagine a real audience and encourage them to become better readers. Bailey (1998) discusses both advantages and disadvantages to holistic scoring. One of the advantages of a holistic scoring system is that high rater reliability can be achieved, and the scoring scale can provide a public standard understood by teachers and students alike.

The third reason for using a holistic scoring system in writing assessment is that when compared to traditional ways of writing assessment, holistic scoring is a more valid system (Schwalm, 1993; Miller, 1997).

On the other hand, the disadvantages of holistic scoring are that the single score may mask differences across individual compositions and does not provide much useful diagnostic feedback to learners and teachers. Furthermore, broad scales may fail to capture important differences across various writing tasks.

Comparing Analytic and Holistic Scoring Systems

White (1989) suggests that the flexibility of holistic scoring allows the scorers to design the kind of test that will meet the needs of their particular students and program. A more recent study by Wolcott (1996) reports that holistic scoring is more efficient than analytic scoring because the student writers focus on composing short essays, and their study of grammar and mechanics is an emphasis on writing as a means of communication and self-discovery.

Hughes has a different view of the comparison of holistic scoring to analytic scoring. Hughes (1997) claims that “The main disadvantage of the analytic method is the time that it takes. Even with practice, scoring will take longer than with the holistic method. Particular circumstances will determine whether the analytic method or the holistic method will be the more economic way of obtaining the required level of scorer reliability” (p.94).

On the other hand a disadvantage of holistic scoring scales may occur when the number of the descriptors increase. A problem may arise at that time as the instructors will be confused to decide on the grade. Another important point about the descriptors is that, when the descriptors become too detailed, instructors will be confused again to decide on the grade that they will assign to the student’s paper (Hughes, 1989).

Results indicate that, when compared to traditional writing assessment, holistic ratings of class work can be done with a high level of rater agreement and the ratings can discriminate among grade level and genre differences in students’ competence (Baker & Linn, 1992).

Problems of Writing Assessment

The problems of writing assessment will be analyzed under the sub-titles of Fairness and Training. What is feasible and useful in one institution may not be elsewhere, but no matter what method of assessment is adopted teachers must be trained in order to assure fairness.

Fairness

Fairness can be defined as the degree to which a test treats every student the same or the degree to which it is impartial. So teachers and testers often do everything in their power to find test questions, administration procedures, scoring methods, and reporting policies so that each student will receive equal and fair treatment (Brown, 1996)

Fairness of the written task relates to the question of time allowed to complete the task, the degree of difficulty at the task, the scope of the topic and elements of bias in the topic. To be fair, the task must obviously set reasonable expectations, considering the amount of time permitted for the test (Jacobs et al.1981).

The impact of writing assessment on students is apparent to any person involved in academic learning contexts. Grabe & Kaplan (1996) argue writing assessment to be a major determinant of students' future academic careers, whether through an in-class assessment of student progress or a standardized proficiency assessment. In the same article Grabe & Kaplan continue as follows:

Writing is commonly used to assess not only students' language skills but also their learning in many academic content-areas. For this reason, among others, the ability to provide students, teachers, administrators with fair and supportable assessment approaches is a serious issue. Not only do many

decisions rest with writing assessment, but assessment processes have a great impact on student attitudes and their motivation for future work (p.377-378).

When we speak of the fairness of writing assessment we are referring to test evaluators; the people who will read the compositions to judge the writers' abilities. If individual evaluators are inconsistent in applying the criteria or standards of the evaluation from paper to paper, or if they differ markedly from other readers in their judgements of the same paper, the scores they assign to the test papers will not be reliable indicators of students' writing proficiency. In such cases, it will not be certain whether the test scores reflect the actual abilities of the writers or are more a function of the readers who evaluated the papers (White, 1989; Grabe & Kaplan, 1996).

Fairness of writing assessment is concerned with many issues. However, the primary concern here is the rater reliability. Jacobs et al. (1981) have categorized the five factors contributing to the reader variable.

The first one is the different standards of severity. Some readers may be relatively lenient as they assess the quality of papers, assigning mostly high scores to all papers; others may be relatively severe, giving scores only from the lower portion of the scale.

The second reader factor may happen if the readers distribute scores differently along the scoring scale. Some readers may assign scores from all points of the scale while others may give scores closer to the middle of the scale. This kind of reader variable is also reported by Brown (1996) who claims that instructors should ensure that they are using all portions of the scale.

Readers' inconsistency in applying the standards of the evaluation is the third factor. Readers who are inconsistent during an evaluation session may award high marks to the papers of a certain quality early in the session, only to give low marks to the same papers or ones of similar quality later in the evaluation session.

The fourth reader factor may happen when readers react to certain elements in the evaluation or in the papers. Certain aspects of the evaluation itself, often unrelated to the quality of student compositions, may shape the readers' perceptions or expectations of the writers before they even look at the test compositions. Similarly, Lloyd-Jones (1987) notes that raters will favor some writers over others in a test or during the assessment of a test because of the nature of our individual approaches to evaluation.

The fifth and the last factor is the readers' value of different aspects of a composition. Readers do not always agree on the qualities that are important in student writing. What may be a critical determinant of excellence for one reader may be relatively unimportant to another. White (1989) also states that sometimes it is hard for the graders to agree on the same criteria to be involved in the evaluation process of holistic scoring system.

The presence of this reader variable is well illustrated in the 1961 study conducted by Diederich, French and Carlton (cited in Jacobs et al., 1981). They asked 53 readers from several occupational backgrounds to evaluate 300 composition papers written by L1 college freshmen. The readers were not given standards or criteria on which to judge the papers. Diederich et al. found that the judgements of the readers varied wildly. Of the 300 papers, every paper received at least five different grades. Surprisingly, 94 percent received seven, eight or nine different

CHAPTER 3: METHODOLOGY

Introduction

The major focus of this study is to identify the reliability of the old analytic writing assessment system used at Anadolu University Preparatory School and to find out if the use of a standard grading format and holistic scoring will result in an increase in the reliability of writing assessment at Anadolu University Preparatory School. In other words the major focus of my study is to understand to what degree is the reader variable at work at Anadolu University Preparatory School and to what extent the use of a holistic scoring system will help the instructors to grade both more reliably, and thus, fairly.

The study was conducted at Anadolu University Preparatory School, and the participants were selected from the instructors of the same institution. This is an correlational study.

In this chapter, the participants involved in the study, the instruments used to collect data and the procedures employed are discussed in detail.

Participants

The participants involved in the study were chosen from Anadolu University Preparatory School writing instructors. Six writing instructors out of eight were chosen to participate in the study. All of the instructors who participated in the study were teaching different levels of the program. Three of these instructors were male; the other three were females. All female instructors had more than 10 years of language teaching experience, while all three male instructors had less than three years of language teaching experience.

The instructors were aged between 23 and 42 and all of them were non-native speakers of English.

Among the six instructors who participated in the study four instructors were teaching writing during the 1998-1999 spring semester. Anadolu University Preparatory School uses a skills based syllabus so the other two female instructors who were not giving writing courses during that semester had given writing courses beforehand.

However, in the final statistical analysis one of the instructors was excluded from the study for being an outlier. So the number of the instructors used in the statistical analysis was five.

Materials

In order to find the inconsistencies which arise from individual teachers' writing assessment practices each instructor was given a file consisting of 15 student papers (See Appendix A). These papers were chosen from the papers written for the September 1998 placement exam at Anadolu University Preparatory School. While choosing the papers I took care to select papers reflecting a range of scores from 35 to 85. There were three papers with 35 points, three papers with 45 points, three papers with 55 points, three papers with 65 points, and three papers with 85 points. The papers that were chosen had already been graded by other instructors at Anadolu University after the placement exam, but the instructors involved in this study were never told the grades that had been given to the papers in order not to affect their judgements of the papers. All the papers were retyped with the original errors made by the student writers. No other changes were made in the papers.

The papers which were used during the training session were also chosen from previously written student papers. The only difference was that the papers used for training were not exam papers; they were from my students' assignments during the 1997-1998 academic year. Among the papers that were available during that time, the low range of papers were missing so I used some papers from Bilkent University freshmen. While choosing the papers for training I was careful about the qualities of the papers again, but this time in a different manner. Sample student papers for each band of the holistic scoring system I had developed were chosen. During the training session the instructors were presented 30 student papers in order to introduce the new writing assessment system. Again all the papers used at the training session were retyped with the errors which had been made by the student writers and no other changes were made in the papers.

Instructors were then trained on a holistic scoring scale which was developed especially for Anadolu University Preparatory School (See Appendix B). The holistic scoring scale consisted of five bands. Each band was defined with a definition sentence and descriptors that related to content, organization, grammar and vocabulary.

Procedures

I began the study by asking Prof. Dr. Gül Durmuşoğlu Köse, Head of the Preparatory School for formal permission to conduct the study in the preparatory school of Anadolu University. The instructors who were participated in the study were not told the focus of the study in order to not to be affected by it.

Upon receiving permission, I chose 15 student papers among the papers of the placement test which was given at the beginning of the 1998-1999 education year.

The next step was to give the same 15 papers to the instructors and ask them grade the papers with the old analytic writing assessment system they use in the institution. According to the old analytic assessment system the instructors were given a scale of 30 points for content, 30 points for organization, and 40 points for grammar. All six instructors were given the papers at the same time, and the papers were collected from the instructors two days later.

The next step was to train the instructors on the new writing assessment system. The training was done at Anadolu University Preparatory School with the five writing instructors who participated in the study. One instructor was trained separately because of a schedule conflict. The instructors were trained on the holistic scoring scale which I had developed for Anadolu University Preparatory School. The training session was held one and a half months after the first grading session, since I did not want the instructors to remember in the post-training test, the grades that they gave to the papers. First, all the instructors were introduced to the holistic scoring scale. Then sample student papers for each scale were introduced to. Thirdly, all the instructors were given sample papers to grade following the holistic scoring scale that they had been introduced. The instructors were told to read each paper twice and grade it without correcting anything they saw on the papers. The instructors were not given exact time to grade the papers. I was in charge of watching them and when I decided that they had finished grading we discussed the grades that they assigned to the papers and their reasons. The main reason for that was to enable the instructors to get used to the scale. The training lasted for about four hours. The training session was tape recorded and it was transcribed.

The same 15 papers that the instructors had graded before the training were then given to the instructors again, and they were told to grade the papers using the holistic scoring scale that they had been trained on. All the instructors were given 2 days to complete the grading of 15 papers.

Data Analysis

When the papers were returned by the six writing instructors who participated in the study, I had two different grades for the same papers given by six different writing instructors.

In order to find the reliability of the two writing assessment systems, a One-way ANOVA was used. After finding the mean scores of each system with a One-way ANOVA, the Student T-Test was used to compare the mean scores of the two writing assessment systems. The results of these tests are reported in the next chapter.

CHAPTER 4: DATA ANALYSIS

Overview of the study

The major focus of this study is to identify the reliability of the old analytic writing assessment system used at Anadolu University Preparatory School and to find out if the use of a holistic scoring system will result in an increase in the reliability of the writing assessment. In order to achieve this purpose, I asked the instructors, who participated in the study, to grade student essays. Each instructor graded the same paper; first with the old analytic writing assessment system used at Anadolu University, second with the new Holistic Grading system which they were trained on. Having two grades for each paper from five different instructors, before and after the training, I had the chance to compare the grades and evaluate whether the use of a standard grading format will result in an increase in the reliability of the writing assessment at Anadolu University.

Participants

The participants involved in the study were chosen from Anadolu University Preparatory School. Six writing instructors, who were teaching different levels were chosen. Three of the instructors were male; the other three were females. All female instructors had more than 10 years of language teaching experience, while all three male instructors had less than three years of language teaching experience. The instructors were aged between 23 and 42. All the instructors involved in the study were non-native speakers of English.

Among the six instructors who participated in the study, four were teaching writing during 1998-1999 spring semester. The other two female instructors who

were not giving writing courses during that semester, had given writing courses previously.

The first step of the study, grading 15 student papers with the old analytic system, was carried out by six instructors, but after the training on the new writing assessment system the number of the graders was reduced to five. Instructor E was excluded from the study because of being an outlier. When the grades that the instructors gave to the papers after the training were analyzed, it was seen that the grades that the instructor E gave were at least 2 or sometimes 3 bands higher than the other instructors. This is the reason why instructor E was excluded from the study.

Materials

Data were collected by means of having each instructor grade student essays. In order to investigate individual instructors' writing assessment practices, each instructor was given a file consisting of 15 student papers (See Appendix A). These papers had been chosen among the papers of the September 1998 Placement Exam which was given at Anadolu University Preparatory School. While choosing the papers, I took care to select a variety of grade levels ranging from 35 to 85 out of 100. The papers had already been graded by other instructors of the institution after the placement exam, but the instructors who participated in the study had never seen these papers, nor been told the grades given to them. In order to make it easy for the instructors to read the papers, all the papers were retyped with the original errors made by the student writers. No other changes were made on the papers.

Instructors were presented with other student papers during the training session. These 30 student papers were only used in order to train the instructors on the new system, and were different from the papers which the instructors were given

to grade. This time the student papers were chosen from students' assignments during the 1997-1998 education year. Also, some of the papers that were presented during the training were taken from Bilkent University Freshman English unit. Of the papers available to me at the time, not enough fell into the lowest bands so these were filled with papers from Bilkent University freshmen unit. Again, the papers were chosen to present a variety of levels of quality. Before introducing the new system to the instructors, I chose sample student papers for each band of the holistic scoring system I had developed (See Appendix B). Student papers having all the characteristics of a 1, 2, 3, 4 or 5 quality paper were presented to the instructors. During the training session, the instructors were not only presented papers but also were trained to use the new system. Instructors were given extra papers to grade for this training. Again, all the papers used at the training session were retyped with the errors which had been made by the student writers and no other changes were made on the papers. Following training, the instructors were given the 15 papers they had scored using the old analytic system and asked to re-score them using the new holistic system.

The new system that the instructors were trained in consisted of five bands. Each band was defined by a definition sentence and descriptors that related to content, organization, grammar and vocabulary.

Data Analysis Procedures

In order to find out the reliability of both systems, first the results of the scoring with each system were analyzed with a One-way ANOVA. With this analysis, each system's sum of squares and mean squares were found. Then, having the mean score of each system, the Student T-Test was used and the paired

differences between the systems -mean, standard deviation, standard error of the mean- were found. The reason for using the Student T-Test was to compare the mean score of each system and see if there was a difference between the two systems.

Results

Before comparing the scores of the papers given by the same instructors before and after the training, I found the mean score of the papers' Placement Exam grades. Those grades were assigned by instructors who were also working at Anadolu University Preparatory School, but they did not participate in the study.



Table 1

Grades Given To The Papers After Placement Exam

No Of Paper	Placement Exam Grades
1	35
2	50
3	65
4	35
5	65
6	85
7	50
8	50
9	85
10	20
11	85
12	75
13	75
14	70
15	65

These student papers were graded by the instructors working at Anadolu University Preparatory School using the following criteria: 30 points for Organization, 30 points for Content, and 40 points for Grammar. The mean of the grades is 60.67.

After finding the mean score of the grades after the placement exam, I calculated the mean score of each paper from the grades given by the five writing

instructors. My aim was to see if by using the same system, inconsistencies between the scores would occur, when papers were graded by different instructors. The mean of the scores given to the papers by five different instructors was found to be 62, which was very close to the mean of the scores given to the papers after the placement exam. However, a close analysis of the means of the individual papers show important differences.

Table 2

Comparison Of The Means Of The Grades Given To Each Paper

No Of Paper	Placement Exam Grade	Means Of Five Instructors
1	35	50.5
2	50	58.3
3	65	71.6
4	35	51.2
5	65	50.2
6	85	64.8
7	50	63.8
8	50	52.6
9	85	71.3
10	20	36.6
11	85	79.5
12	75	75.3
13	75	48.6
14	70	84.5
15	65	71.2

The scores from Table 2, above, were given based on the same criteria and the same papers as those in Table 1. Only the graders are different. This can take us to the issue of whether the same instructors working for the same institution assign different grades to the same papers using the same assessment system provided for assessing writing.

Among the 15 papers only paper 12 received the same grade from different graders with the same system. The score of paper number 8, is nearly the same after the two assessments revealing a 2.6 points difference between the two assessments. A discrepancy is apparent in the paper number 6. The grade given to the paper after the placement exam is 85, but when the same paper is graded by different instructors with the same assessment system, the grade assigned to the paper becomes 64.8. This means that there is a difference of nearly 20 points between the grades.

The results show that some of the grades given to some of the papers increase, whereas some of the grades decrease. We can see that the grades of the papers after the placement exam are the same, but when the same papers were graded by five writing instructors, the mean for paper number 12 was 75.3, which is similar to the placement exam grade. But, when paper number 13 is analyzed, we see that the mean grade is 48.6, this is a dramatic difference of almost 27 points, an undesirable situation for any institution. These differences in scoring may be because of the number of the graders. The greater the number of the graders, the more the differences can occur without a standard grading format (Jacobs et al., 1981).

Another reason for the differences in scoring may relate to the perception of the scoring criteria by different graders. Although, all the instructors were told to

give 30 points each to content and organization and 40 points to grammar both for the placement exam and for the study, it was not clear what “30” meant. Jacobs et al. (1981) call this a “Reader Variable” and describe it as follows: “The variety of readers’ value of different aspects of a composition is a kind of reader variable. Readers do not always agree on the qualities that are important in student writing. What may be a critical determinant for one reader may be relatively unimportant to another” (p.26).

The result is that even though a relative value for each criterion was already determined, there was no guarantee that the instructors attended to those relative values.

Table 3 gives the grades of the instructors that they have assigned to the student papers using the old analytic scoring system.

Table 3

Instructors' Grades Given To 15 Papers

No Of Paper	Instructor A	Instructor B	Instructor C	Instructor D	Instructor F
1	52	40	50	60	35
2	63	65	58	55	30
3	60	80	87	65	65
4	47	50	50	60	40
5	33	65	55	45	50
6	37	65	80	75	70
7	29	85	59	70	75
8	54	50	67	40	30
9	83	70	75	70	60
10	28	40	37	40	20
11	71	90	73	65	92
12	75	90	83	45	85
13	42	40	52	65	25
14	89	95	93	60	95
15	66	85	84	60	65

Appendix C shows the grades for content, organization and grammar that the instructors assigned to the 15 papers with the same criteria that was used to grade the papers after the placement exam. Here we can see differences between the grades given for each of the criteria on different papers.

With this system each instructor is free to grade the paper in terms of grammar from 0 to 40. The criteria from 0 to 40 are not specified so writing instructors have no guidance in assessing a grade in each category.

Table 4

The Grades Assigned To Paper Number 4 By All Instructors

	Instructor A	Instructor B	Instructor C	Instructor D	Instructor F
Grammar	15	25	35	10	15
Content	17	10	0	25	15
Organization	15	25	15	25	10

For the paper number 4, instructor C assigned 35 for grammar whereas the instructor D assigned 10 for the same student writer's paper. This means there is a difference of 25 points between two graders' scores to the same paper. Another difference occurs for the content of the same paper. Instructor D assigned 25 for the content of the paper 4 whereas instructor C gave 0 for the same criteria. Assigning 0 for the content of a paper is not defined so it is not clear why this instructor assigned 0.

Table 5

The Grades Assigned To Paper Number 7 By All Instructors

	Instructor A	Instructor B	Instructor C	Instructor D	Instructor F
Grammar	11	30	39	20	25
Content	08	30	05	25	25
Organization	10	25	15	25	25

Another difference can be seen for the paper number 7. When we look at the paper number 7, we can see that the instructor A gave the paper's grammar a score of 11, whereas the instructor C gave it a score of 39. This gives us a difference of 28

points to the same paper by two different instructors. Again, for the content of the same paper the instructor B gave a score of 30 but another participant of the study who is instructor C gave a score of 5 to the same paper. The difference is 25 points this time for the content of the same paper.

To compare the five different grades for 15 different papers by five different instructors, I calculated the range of the grades in order to see how large or small the difference runs between the grades. Table 6 shows the range of grades for each paper given by the participants of the study.

Table 6

The Range Of The Grades

No Of Paper	Low - High	Range
1	35-60	25
2	30-65	35
3	60-87	27
4	40-60	20
5	33-65	32
6	37-80	43
7	29-85	56
8	30-67	37
9	60-83	23
10	20-40	20
11	65-90	37
12	45-90	45
13	25-65	40
14	60-95	35
15	60-85	25

The most significant range among the 15 papers belongs to paper number 7 because one of the instructors graded the paper as 29 and another instructor graded the same paper as 85. This means that the range of the grades is 56 points. This is a substantial difference. If the paper had been graded by the first instructor after the placement exam, then the student would have failed the test and been placed in prep school. If the paper had been graded by the second instructor after the placement exam the student would have passed the exam and started the Freshmen Program. The paper was the same, the student was the same, but there were two different results. If the evaluators are inconsistent in applying the criteria or if there is no clearly set criteria or standards, as in the example, the scores that the instructors assign to the papers will not be valid indicators of students' writing proficiency. So it will not be certain whether the scores reflect the actual abilities of the writers. The student who wrote essay 7 could both have either been assigned to study in prep class for one year or have been placed in his/her department. The student's level of English would be the same, but because of the two evaluators' different grades, s/he would have been placed differently.

The same issue was discussed in Brown (1996). Brown states that placement decisions usually have the goal of grouping together students of similar ability levels and concludes that:

Teachers benefit from placement decisions because their classes contain students with relatively homogeneous ability levels. As a result, teachers can focus on the problems and learning points appropriate for that level of student. To that end, placement tests are designed to help decide what each student's appropriate level will be within a specific program, skill area, or course (p.11).

Paper number 7 is not the only example. It is also possible to see some other differences for other papers too. For example, paper number 12 has a range of 45.

Other examples may be papers 2, 6, 8 and 13. The range of the paper number 2 is 35 points, the range of the paper number 6 is 43 points, the range of the paper number 8 is 37 points and the range of the paper number 13 is 40 points.

All of these differences between the grades point to a lack of consistency between the graders.

This new assessment system is a five band holistic scoring system. In order to explain to the instructors what 1, 2, 3, 4 and 5 mean, I transformed each band to an equivalent 100 point system. With this system;

1 = 53

2 = 63 and above

3 = 73 and above

4 = 83 and above

5 = 93 and above.

The first system that the instructors were told to use to grade the papers was out of 100, but the new system that the instructors were trained on was a five scale scoring system. Table 7 shows the grades that the instructors who participated in the study assigned to the papers with the new holistic scoring system that they were trained on.

Table 7

The Grades Given To The Same Papers By The Same Instructors After The Training On The Holistic Scoring System

No Of Paper	Instructor A	Instructor B	Instructor C	Instructor D	Instructor F
1	2	2	2	1	2
2	2	2	1	1	2
3	2	2	2	2	2
4	1	1	1	1	2
5	1	1	1	1	2
6	2	2	2	2	3
7	2	3	3	3	3
8	1	1	1	1	2
9	2	2	2	2	2
10	1	1	1	1	1
11	3	3	3	2	3
12	2	3	1	1	2
13	2	3	2	1	2
14	2	4	3	3	3
15	3	3	3	3	3

When the grades given to the papers with the new holistic scoring system were analyzed, it was easy to see that there were consistencies between the grades and the graders. The grades given to the papers 3, 9, 10 and 15 were the same by all the instructors who participated in the study. There was only one band difference between the grades given to the papers number 1, 2, 4, 5, 6, 7, 8 and 11. For example, paper number 1 was graded as 2 by the instructors A, B, C and F. Only instructor D assigned 1 to the same paper. Also, there were some papers which the

instructors assigned with grades with a difference of more than 1 band. These papers were 12, 13 and 14. When we look at paper number 14, we see that the instructor A assigned 2 to the paper, however, instructor B assigned 4 to the same paper.

In order to find out the mean scores of each system, a One-way ANOVA was used. Table 6 and Table 7 show the results of the mean scores of the systems after the One-way ANOVA.

Table 8

One-way ANOVA For The Holistic Scoring System

		Sum of Squares	df	Mean Square	F	Significance
VAR00001	Between Groups	3.813	4	.953	1.584	.188
	Within Groups	42.133	70	.602		
	Total	45.947	74			

The result indicates that there is significant relationship between the grades given to the same paper by five different graders after the training. There are only 5 chances in 100 that the differences between the grades would have occurred by chance. We can see that the degrees of freedom is 4 and it is significant at the level of .05.

Table 9

One-way ANOVA For The Old Analytic System

		Sum Of Squares	df	Mean Square	F	Significance
VAR00001	Between Groups	2117.387	4	529.347	1.479	.218
	Within Groups	25053.7	70	357.910		
	Total	27171.1	74			

The result for the old analytic system used at Anadolu University Preparatory School is the same as the holistic scoring system. The result indicates that there is significant relationship between the grades given to the same paper by five different graders with the old analytic system used at Anadolu University. There are only 5 chances in 100 that the differences between the grades would have occurred by chance. We can see that the degrees of freedom is 4 and it is significant at the level of .05.

After finding the results of the One-way ANOVA for each system, the Student T-Test was used in order to compare the mean scores of two systems.

Table 10

Paired Samples Statistics

		Mean	N	Standard Deviation	Standard Error Mean
Pair 1	A	3.0360	75	.9581	.1106
	VAR00004	1.9733	75	.7880	9.1E-02

Table 11

Paired Samples Correlations

		N	Correlation	Significance
Pair 1	A & VAR00004	75	.526	.000

The results of the paired samples statistics and paired samples correlations indicate that there is a large difference between the two writing assessment systems used at Anadolu University Preparatory School. The significance is .000 which means there is large difference between the grades given to the same papers by using two different systems.

Table 12

Paired Samples Test

Paired Differences								
95% Confidence Interval of the Difference								
		Mean	Std Deviation	Std Error Mean	Lower	Upper	t	Sig (2-tailed)
Pair1	A- VAR00004	1.0627	.8631	1.0E-01	.8641	1.2613	10.662	.000

The results of the Student T-Test indicate that there is a significant difference between the scores given to the same papers by two different systems. There are only 5 chances in 100 that the differences in mean scores between the two scoring systems have occurred by chance.

An important point to consider is that the results of the Student T-Test only show us if there is a difference between the scores given to the same papers by two different systems. Since we did not have a control group we cannot say one system is better than the other. In order to reach to a conclusion like “System A is better than System B” there should be a control group or a system which has been accepted by everybody.

The Student T-Test was used to compare the mean scores of two groups.

Since the old analytic writing assessment system used at Anadolu University Preparatory School was scaled to 100, and the holistic scoring system was a five point scale, the mean of the old analytic system and the mean of the new system were impossible to compare. Therefore, I divided the scores of the old analytic system into 20. In this way I managed to approximate the two different systems' scores. Table 13 shows this comparison of means.

Table 13

Scores Given To The Same Papers Before And After The Training

No Of Paper	Instructor A		Instructor B		Instructor C		Instructor D		Instructor F	
	B	A	B	A	B	A	B	A	B	A
1	2.6	2	2	2	2.5	2	3	1	1.8	2
2	3.2	2	3.3	2	2.9	1	2.3	1	1.5	2
3	3	2	4	2	4.4	2	3.3	2	3.3	2
4	2.4	1	2.5	1	2.5	1	3	1	2	2
5	1.7	1	3.3	1	2.3	1	2.3	1	2.5	2
6	1.9	2	3.3	2	4	2	3.8	2	3.5	3
7	1.5	2	4.3	3	2.9	3	3.5	3	3.8	3
8	2.7	1	2.5	1	3.4	1	2	1	1.5	2
9	4.2	2	3.5	2	3.8	2	3.5	2	3	2
10	1.4	1	2	1	1.9	1	2	1	1	1
11	3.6	3	4.5	3	3.7	3	3.3	2	4.6	3
12	3.8	2	4.5	3	4.2	1	2.3	1	4.3	2
13	2.1	2	2	3	2.6	2	3.3	1	1.3	2
14	4.5	2	4.8	4	4.7	3	3	3	4.8	3
15	3.7	3	4.3	3	4.2	3	3	3	3.3	3

Note: B means the scores given to the papers before the training

A means the scores given to the papers after the training

When we analyze the two scores given to the same papers before and after the training, we can see that the scores in the new scale are lower in general. Among the 75 scores only the scores of four papers are higher. These are the score given to the paper number 7 by instructor A, the score given to the paper number 7 by instructor C, and the scores given to the papers 1 and 2 by instructor F.

Among the 75 scores the scores of only five papers remain the same; 1, 4, 10, 14 and 15. Instructor B graded the paper number 1 both before and after the training as 2. Instructor D graded the same papers number 14 and 15 both before and after the training as 3. Lastly, instructor F graded the paper number 4 before and after the training as 2 and paper number 10 before and after the training as 1.

Another important result seen in Table 13 is that although the instructors were given papers selected from different levels of quality to grade, none of them graded a paper a score of 5, only one instructor gave one paper a score of 4. This means the instructors did not use the whole range of the scale. The new system had been developed that the instructors should have made use of the whole range of the scale; the fact that they did not, suggests that the new scoring system needs to be reworked.

CHAPTER 5

Overview of the study

This study investigated the reliability of the analytic writing assessment system used at Anadolu University Preparatory School and asked if the use of a holistic scoring system would result in an increase in the reliability of the writing assessment system. In order to achieve this purpose, six writing instructors, teaching to different levels, participated in the study. Instructors were given a file consisting of 15 student papers which had been selected from the institution's 1998 placement exam. Instructors were told to grade the papers with the old analytic writing assessment system used at Anadolu University Preparatory School. After 1.5 months the same instructors were trained on the new scoring system which was developed for this school. Instructors were presented the new system by using papers chosen from student assignments. After the training the instructors were given the same 15 papers that they had already graded with the old analytic system. Instructors were then told to grade the papers with the new holistic scoring system.

Having the grades given to the same papers by the same instructors with two different systems, the data were analyzed first by using a One-way ANOVA in order to find the mean scores of each system. Second, the Student T-Test was used to see the differences between the mean scores of the two systems.

Summary of Results and Conclusions

This section discusses the findings and the conclusions that have been drawn through the process of data collection in order to answer my research questions. Each

sub-section refers to one research question. In addition, where relevant, some of the studies that have been reviewed in the literature are referred to.

The Reliability of the Old analytic Writing Assessment System Used at Anadolu University Preparatory School

The grades given to the same papers by five different instructors with the old analytic writing assessment system used at Anadolu University Preparatory School showed that there occurred inconsistencies across the papers and the instructors (See Table 3).

The inconsistencies in the instructors' grades occurred for all papers. Although a few papers were graded by two instructors with the same score, in general, all five scores were different from each other for each paper.

Then the range of the five scores for each paper was found (See Table 6). The table shows that the most significant paper with the greatest variation is paper number 7 with a range of 56, which means there are 56 points difference between the lowest and highest scores that any two instructors assigned for that same paper.

This leads us to the conclusion, like that of White (1989), that without a clearly stated scoring criteria each paper can get every possible score from different scorers.

In order to find the reliability of the old analytic writing assessment system used at Anadolu University Preparatory School, a One-way ANOVA was used (See Table 9). Unexpectedly, the results indicate that there is significant relationship between the grades given to the same paper by the five different instructors with the old analytic writing assessment system at the level $p < .05$. This means that there are

only 5 chances in a 100 that the differences between the grades would have occurred by chance.

Although the statistical analysis shows that the old analytic system is reliable within itself, this does not explain the big differences given to the same papers by different instructors. As was stated in Brumfit's study (1993), one of the purposes of assessment is whether the students have learnt all that was required by the syllabus, but here the situation is different: one student may be thought to have learnt the subject by an instructor and may be given a high score, but the same student may be thought to be unsuccessful in learning the subject and may be given a low score by another instructor.

This is an unwanted situation both for the students and the instructors because instructors should understand the weaknesses and strengths of students after the assessment and try to help students where they need particular help.

In the same study Brumfit (1993) states another important purpose of assessment is to understand whether the instructor uses a successful methodology. After the assessment, being sure that it is both reliable and valid, the instructor should also assess the method that s/he used to teach the subject. Under those circumstances, having different grades from different instructors, the instructors will not be sure whether the method that they used worked or not.

The Reliability of the New Holistic Scoring System Used at Anadolu University Preparatory School

Again using a One-way ANOVA for the new system, I found the reliability of the new system within itself. The results indicate that there is significant relation between the grades given to the same paper by five different graders after the

training at the level of $p < .05$. There are only 5 chances in a 100 that the differences between the grades have occurred by chance.

This is consistent with Baker & Linn's (1992) study, in which it is noted that, holistic ratings of classwork can be achieved with a high level of rater agreement and the ratings can discriminate among grade level and genre differences in students' competence.

It is also related with what Miller (1997) has pointed out in his study. According to Miller, one of the advantages of holistic scoring is that it can be a highly reliable method of grading.

The grades that the instructors gave to the papers with the new holistic scoring system also show that the instructors were using the low portions of the scale. Although the papers were chosen especially presenting all levels of quality, instructors did not assign grades from the high portions of the new holistic scoring system. The grades were usually between 1, 2 and 3. None of the instructors gave the score of 5 to any of the papers, although there were 5 quality papers among the ones that they were given to grade with the new holistic scoring system.

In order to solve that problem, more training on the new holistic scoring system or revision of the scale can be suggested.

The Differences Between the Grades Given to the Papers in the Two Different Scoring Systems

After finding the mean scores of two different writing assessment systems by using a One-way ANOVA, I used the Student T-Test in order to compare the reliability of the two systems (See Table 12).

The Student T-Test results show that there is a significant difference between the scores given to the same papers by two different systems at the level $p < .05$. There are only 5 chances in a 100 that the differences in mean scores between the two scoring systems have occurred by chance.

Limitations of the study

My study has certain limitations regarding participants and statistical analysis which should be mentioned. I was planning to use six writing instructors, three of them native speakers of English. In this way I could also mention whether there was a difference between native and nonnative instructors in grading a writing paper. However the three native instructors whom I approached for help in the study refused to participate. I consider this a limitation to my study because I missed the chance to see the difference between native and nonnative instructors.

Another limitation concerns one of the instructors involved in the study. Unfortunately, one of the instructors who participated to the study was excluded because of being an outlier in the second round of grading. Although s/he participated in the training, somehow all the grades that s/he gave were at least 1 or 2 bands higher than the other instructors. This might be because of different standards of severity. Some readers may generally use the higher portion of the scale whereas some may use the lower portions of the scale (Jacobs et al., 1981).

Another limitation is that one of the instructors could not attend the training because she was abroad. An individual training session was held for her when she returned back. The process and the program of the training was the same. She was trained exactly in the same way that the other instructors were trained. However, it would be preferable for all participants in similar studies to be trained together.

Another limitation of my study was about the number of the papers that the instructors were given to grade. Using 15 student papers may seem to be enough, but if the number had been able to be increased, I believe more valuable information could have been collected. The reason why I used only 15 student papers is that writing instructors were already busy with reading their own students' assignments and exams so having them read more than 15 papers would be extra work. Also, I thought that if they were given a number of papers that they could not handle this would lower the quality of the work that they would do.

Because of lack of a control group we cannot tell which system is superior. We can only tell that there is a difference between the grades given to the papers by two different systems.

Curricular Implications

The results presented in detail imply a few implications that should be taken into consideration. In this part, each is discussed in detail.

Scoring Criteria

The holistic scoring system that was developed for Anadolu University Preparatory School was originally developed for this study. I have been working at Anadolu University for three years and I know the backgrounds, needs and expectations of the students. Combining these with the objectives of the institution I developed the new holistic scoring system.

When the writing instructors were presented the new scoring system during the training, we had time to discuss the deficiencies of the new system. Instructors reported that it would take time to get used to the new system. Instructors also reported that the scores given to each band were not appropriate to our scoring

system at Anadolu University Preparatory School. As a solution, one of the instructors suggested that me change the new system into an analytic holistic scoring system. As a next step, not only the writing instructors who participated in the study, but also the other writing instructors working at Anadolu University Preparatory School may come together and discuss the advantages and disadvantages of this new scoring system and decide on the most useful and appropriate scoring system for the institution, instructors and students.

Another implication about the scoring criteria may be about the descriptors of each band. After the training session the instructors reported that they sometimes had difficulty deciding on a grade. They believed that some of the descriptors of some bands were similar to each other so they could not decide on a grade easily. In order to avoid this problem, the descriptors of each band may be examined and evaluated further.

Effects on Instructors

As was stated above only six of the writing instructors working at Anadolu University Preparatory School participated in the study. As there are many other writing instructors working at the institution these instructors should also be presented the new writing assessment system developed for the institution and trained on it.

One of the most important pedagogical implications for the instructors is that the use of a standard grading format should result in a unity between the instructors and also the grades that those instructors assigned. It should not be forgotten that when there is inconsistency among the instructors working for the same institution the validity of the assessment system used at the institution can be questioned. It is

hoped that the use of the new holistic writing assessment will omit the inconsistencies among the instructors and will help develop unity between the instructors.

The instructors who participated in the training session raised other positive effects of the new system. They reported that it took less time for them to grade the papers with the new holistic scoring system. This is related to what Schwalm (1993), Brown (1996), and Williams (1996) have reported. Although the instructors were not used to the new system it was quicker for them to assign a grade to the papers. Being a quick way of writing assessment system was mentioned in Brown (1996) as, if the instructors using the holistic scoring system are experienced they will grade a paper in less than a minute.

One of the instructors also reported that she had felt as if she was reading the diary of a student rather than an essay which was written to be graded. This is also consistent with Miller's (1997) study. Miller stated that holistic grading prevents the reader from concentrating only on grading, and, as a result the instructor reads student work more naturally, as if she were reading any text.

Training

In order to present the new writing assessment system, the instructors who participated in the study were trained on the new system. The training session took a whole afternoon, nearly four hours.

As was stated in the limitations of the study one of the instructors was excluded from the study because of being an outsider although s/he participated in the training. This brings the conclusion that the training was not useful for that

instructor. Supposing that the same problem may occur in the future the training session program may be reorganized to avoid such problems.

The student papers that the instructors were given to grade were especially chosen to provide models in each band. When the grades were analyzed I saw that all the instructors had predominantly used the lower portion of the scale. Although there were 5 quality papers none of the instructors gave a score of 5. Only one instructor gave a score of 4 to a paper. The grades were generally 1, 2 or 3. This means that the instructors used the lower portions of the scale. This leads us to the conclusion that this training session was not adequate for the instructors to make proper use of the scale; they may require more training on the new writing assessment system. Another conclusion may be that the rating scale must be adjusted with the help of the instructors who participated in the study and the instructors giving writing courses.

This study set out to examine the reliability of the old analytic and the new holistic scoring systems for writing assessment at Anadolu University Preparatory School. Statistical analysis showed that both systems were reliable within themselves. However, the superiority of either system could not be determined because of a flaw in the design of the study. While both systems were found reliable, qualitative analysis showed that problems remain in both systems.

REFERENCES

- Alderson, J.C. & Clapham, C & Wall, D. (1995). Language test construction and evaluation. New York: Cambridge University Press.
- Bailey, Kathleen M. (1998). Learning about language assessment: Dilemmas, decisions, and directions. Pacific Grove: Heinle & Heinle Publishers.
- Baker, E. & Linn, R. L. (1992). Writing for portfolios: Potential for Large Scale Assessment. California: Apple Computer, Inc.
- Bates, L. & Lane, J. & Lange, E. (1993). Writing clearly: Responding to ESL compositions. Developing grading standards. pp:103-130. Boston: Heinle & Heinle Publishers.
- Brown, J. D. (1995). The elements of language curriculum: A systematic approach to program development. Boston: Heinle & Heinle Publishers.
- Brown, J. D. (1996). Testing in language programs. New Jersey: Prentice-Hall, Inc.
- Brumfit, C. (1993). Testing literature. In C. Brumfit, (Ed.), Assessment in literature teaching. Hong Kong: Macmillan Publishers Lim.
- Bullock, R. (1991). Autonomy and community in the evaluation of writing. In R. Bullock & J. Trimbur, (Eds.), The politics of writing instruction: Postsecondary. Bynton: Cook Publishers, Inc.
- Dalton, E. (1993). For many students and teachers "Test" is a detestable word. Available: <http://www.ballard-tighe.com/ideas/idea.htm>.
- Dinan, J. S. & Schiller, S. A. (1995). Using essays for outcomes assessment. Available: <http://www.ehhs.cmich.edu/ins/dinan>.

Fulcher, G. (1996). Assessing writing. Writing in the English Classroom, 6 (2), 91-107.

Grabe, W. & Kaplan, R. B. (1996). Theory and practice of writing. New York: Addison Wesley Longman Limited.

Hughes, A. (1989). Testing for language teachers. Glasgow: Cambridge University Press.

Jacobs, H. L. Zinkgraf, S.A., Wormuth, D.R., Hartfiel, V. F. & Hughey, J. B. (1981). Testing ESL composition : A practical approach. Rowley : Newbury House Publishers, Inc.

Lloyd-Jones, R. (1987). Tests of writing ability. In G. Tate (Ed.), Teaching composition:12 Bibliographical essays. Texas: Texas Christian University Press.

Madsen, H.S. (1983). Techniques in testing. Hong Kong: Oxford University Press.

Miller, M. (1997). Alternative grading models: Holistic grading. Available: <http://www.cstw.ohio-state.edu/tutor/holist.htm>.

Moursund, D. (1995-1997). Available: <http://www.iste.org/Resources/Projects/Road Ahead/assess.html>

Schwalm, K. (1993). Holistic scoring. Available: <http://www.mcli.dist.maricopa.edu>.

University of Texas at Austin. (1997). Instruction and assessment. Available: <http://www.tnet.edu/teks/science/stacks/instruct/scsubjec.html>.

White, E. M. (1989). Defining purposes and issues in the assessment of writing ability. Chapter 5. pp: 89-112. Developing successful college writing programs. California & London : Jossey- Bass Inc., Publishers.

Williams, D. J. (1996). Assessing writing . Chapter 11. pp: 255-299.

Preparing to teach writing. Belmont : Lawrence Erlbaum Associates.

Wolcott, W. (1996). Evaluating a basic writing program. Journal of Basic Writing, 15 (1), 57-69.

Yancey, K. B. (1997). Writing assessment: A position statement from CCC.

Available: <http://www.sfasu.edu/writing/xwp/assess.htm>.



Appendix A

Student Papers

NUMBER:1

At first, private life must be explained to understand the subject and to have a first idea about this. For me private life is some events, conversations, etc which the person doesn't want to be known by other people. Because of this reason the explanation of private life is very different or has many differences mind to mind. So everybody must be careful when talking about somebody or something.

If we ask "Why private life must be the secret of the people?" This is personal selection. Everybody must have secrets or have secrets. From this reason there is a subject to write a composition about "Private Life".

In my idea if people doesn't have private life their action would be serious or shyness, also this is right for me and I think for everybody except a few people.

Private life of famous people should not be made public. There is 2 reason for this. subject. The first one is they are also people and must have freedom. If a person is in the jewel he can't have an enjoyment in time of he lives. I say jewel but there are many people free but feeling in the jewel. So private life of famous people should not be made public except some private causes like Bill Clinton. I can't say many ideas about this but the judge will say the last words.

The second reason is famous people's strange life may make big characteristic problems on children and adult. If an adult watch, or see or read their strange life they would make everything to have much money which can give them the life they want. If these people earn money from regular ways this would be

good. But I think that ten people in hundred (%10) they will have irregular jobs. or If they can't earn money like they have planned they will smoke cigarette at first and alchol and drugs and a person who is harmful to his public.

Like I said there can be very different ideas about this. Lastly everybody must have freedom. If anything get your freedom it mustn't be regular.



NUMBER:2

I think that the private-life of famous people should not be made public because all people have got the right of life whatever they want.

No matter who you are, you are able to do what you want but you are famous person it seems to be impossible, public is curious about your life what you do, who is your darling, how much you earn etc. They always want to learn your life. Everybody watches your each step and you make wrong behave, they criticize you immediately. if you are top, you are fallen on the top. Media both make you, famous and successful person and investigate your former life and former wrongs. if media find out your former bad behave, media publicate it all people.

For example; Bill Clinton. In my opinion he is successful president. He did not want to explain his private life. it was lived and finished it should not interest us to learn what he does in his private life. if you are not famous people . you can live what you want to but we do not give the live of right to him. He is a people like us . and it should not be important what he does in his private life. His success and his contribute to his country should be important so I don't accuse of him. He is democrat and all he wants is contribute his country and give his public good condition life.

Although you want to live freely, public's curiosity does not reduce. if people's curiosity does not reduce, we are not able to live what we want to. All this reason is curiosity. Much curiosity is danger of our health.

NUMBER: 3

PRIVATE-LIFE OF FAMOUS PEOPLE

Famous people should take a lot care on everything. They don't do everything they want. Because people always wonder about them.

If this kind of people don't take much care on their private-life, they can lose their jobs, family etc. They must always be careful. For example; I can go to the cinema, shout in the street or, maybe, have a lot of girl-friends. But they don't do these things easily. People always want to get some news about them. so that there are a lot of journalists around them to take some photographs or report. Some journalists may want to take photos of them in secret ways. If this kind of people are married and have a sexual correspond with others, just like Bill Clinton, it is not good for them. Because journalists are always on their ways. This time is end of their carriers.

On the other hand, however, if an ordinary man had done the same things like Clinton, the results would have been very different. Public and journalists don't even wonder about him.

Finally, the people that should made public take a lot of care on their private-lives. Unless, they don't do this, they may lose their carriers or everything they own.

NUMBER:4

DO YOU WANT TO BE FAMOUS?

You have a dream, always want to be singer and tell it to the people. But they laugh to you. Because it can't be possible. After years, you are very famous person. All the people wants to have a photo from you. And where you go they always follow you. But for me it isn't good.

Because I can't have a private-life. I must know how I attitude to the people, what I say or before everything what I have to think. For example I don't like some people who often have a life like their thoughts. When I talk with them I ought to say "I love you". But it isn't true. Because what you are you must be it. Otherwise after a time you can lose yourself and what you are you forget.

If you ask me " Mr. Clinton's attitudes are tru or false?" I think everybody is sure that is not true. We are human ofcourse we can do wrong. But if you are famouse and before years you worked for it very much, you must choice work or private-life. Because all of them can't be in the same time. Expecially for me Mr. Clinton's famous isn't important. The important thing is he is married and have a daughter. Before he had a girlfriend he should have thought But he didn't do it. And now we can read his feelings from his eyes. I think nowadays he has always said yourself why I have been the president of the U.S.A. He must ashame first from his family, second from the American people.

Love is always love If you look to the flowers or the animals or the similing people, you can see. Now Mr. clinton has everything (rich, family, jop, a good healty) but by the way he has nothing. And it is the worst thing in the world.

I nevr want to be not only Mr, clinton but also his daughter or his wife. The falses can be forgotton but not all of them. Sometimes we can paid the conclusions with our life.



NUMBER:5

THE SCANDALS

I think these kind of events must be named scandal. But it changes for artists, actors or singers. They can do whatever they want. Because it's their own special life. In addition they must be careful. They mustn't disturb other people's lives.

And It's right for them keeping their special life in secret.

But if these events become in men's of government lifes the problem begins. These people are in front of the eyes of nation. Their special life must be better than the other people. But nowhere in the earth this can not be done.

In my opinion Bill Clinton is only one of them. Otehrs cannot be known. I wish Bill Clinton's life had been in secret. But from now on I can only think poor man.

Apart from these I said the government men mustn't do these scandals. But it's their life. So we can'y say they must be robots. In spite of this opinion, my opinion is different the cases like Bill Clinton's aren't right for me. I said like that because I'm a Turkish person and lots of Turkey's people think like me.

In conclusion I'm repeating: it's their life and it doesn't matter me. My family's and my life are important for me.

NUMBER:6

THE PRIVATE LIFE OF THEM

There are lots of people in the world. Some of them are very important for us. We all want to know about them. Why?

Special people are always with us. However we don't see them we know their lifestyle. We always wonder about what they are doing now, where they are going, how their private-lives are. When I think about wheather their private-lives are important or not I decide this it's importing. Because these people always attract the attantions all of the people. OK! Of course they will have their own private-life but if they are same example to others, it can be made public. In fact some of the ordinary people interested in their social life. That's the true one. I can take lots of famous people for example singers, politicians, actrists ext. In Turkey Tarkan, who is a famous pop singer, is very important for us. His ideas, behaviors, wearing styles are followed by our youngers. One person can change a socials life-style.

Bill Clinton is a famous politician in the world. When we learn his some behaviour of private-life we feel ourself to think how such a famous people live. In addition to this there are youngers or followers whom want to be in his place.

When I think all of them, I can get an idea about this subject. Consequently I can say this "Yes, famous peoples' private-lives are important and it can be made public."

NUMBER:7

I think that the private-life of famous people shouldn't be made public. If a person is a famous people it's not need be just made public his/her private-life.

Nowadays people always think the other peoples life especially private life of famous people. I think human beings are very curious. And also I don't think the famous people are asked whether their private-life should be made public or not. Perhaps some don't object this situation. In my opinion one famous person's life can be made public, but private-life can't be done. These people (famous people) may be unpleasant when they saw their private-life's writing on the newspaper or the other things like this. I completely object to be made public famous people's private life. Not only the olders but also youngers learn the famous people's private life and I don't think it's a useful thing for youngers.

I am curious that what brings on their hands, learning or making public famous people's private-life. How we can live and do the things freely, famous people must do this, too. Every time famous people mustn't be doubtful if they do the wrong thing and it can be made puplic. Everybody must live independence.

Even today, when we are listening the news on television, after some important news, the speaker just began to say which famous people are on holiday, who get married, like these. Are these important things that must be public on television? Almost all people think the famous people's private-life but they don't think today's problems....

NUMBER:8

BEING RESPECTFUL

This subject is almost the number one of the discussing topics. But I think that private life of famous people shouldn't be made public, however it is not impossible. Therefore, we must be calm. For the ordinary people it isn't a problem. The more one person is on eye, it's the more difficult to keep the private-life. Because everyone has a curiosity about their life. For example Bill Clinton. In my opinion Bill Clinton's private life shouldn't be made public. Because his position is extraordinary. If the life of president can be shown like that, the mood health of the people in that country can be deteriorated. On the other hand, everything must be shown for the justice. For the order it's must be done. The one person's acts is conjunctions about other people acts. so everyone must be careful. For the sake of the new generations we must be so respectful about each other's life. We must give them an opportunity to have a life in which doesn't consist stress, fear or other bad things.

NUMBER:9

WHAT HAPPENS IF THE PRIVATE LIFE OF FAMOUS PEOPLE IS MADE
PUBLIC

I think that the private-life of famous people should be made public. If this is done, you will see that there will be some advantages for the society of these famous people live in.

If this kind of people don't take much care on their private-life, they can lost their jobs, family etc. They must always be careful. For example; I can go to cinema, shout in the street or, maybe, have a lot of girl-friends. But they don't do these things easily. People always want to get some news about them. so that there are a lot of journalists around them to take some photographs or report. Some journalists may want to take photos of them in secret ways. If this kind of people are married and have a sexual corresponde with others, just like Bill Clinton, it is not good for them. Because journalists are always on their ways. This time is end of their carriers.

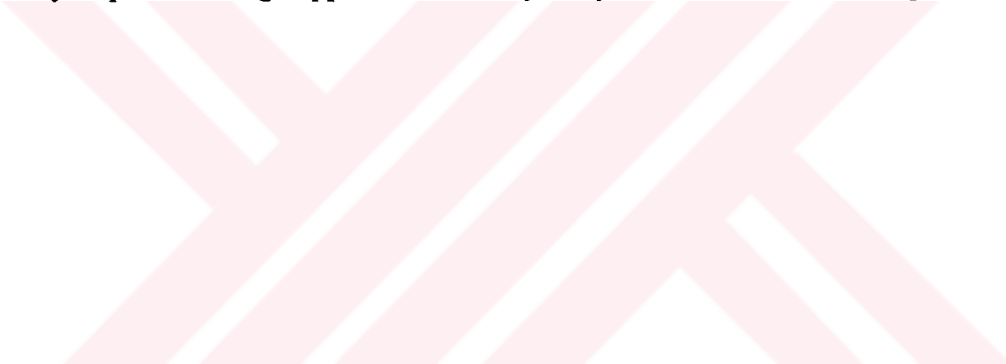
On the other hand, however, if an ordinary man had done the same things like Clinton, the results would have been very different. Public and journalists don't even wonder about him.

Finally, the people that should made public take a lot of care on their private-lives. Unless, they don't do this, they may lost their carriers or everything they own.

NUMBER:10

When I was a child I was spending my time just the things that I could have fun. Nothing was really mattering to me. If I could could have a chance I would like to be a child forever. But unfortunately year by year I started to become an adult. Parallel to this I started to interest in happenings of real world. It wasn't funny as the toy world but it was interesting. In these days I began to watch the evening news. The speaker was telling about the things which have a relationship with my life.

A few weeks ago when I was watching the news I heard the voice of the speaker who is saying "the big scandal" with the big letters. I was excited. I thought a very important thing happened that every body had to learn something about it.



NUMBER:11

WHERE TO?

Nowadays, world's turning so fast that we aren't able to catch some important points which are necessary. When we miss them, we begin to spend time about nonsense events that we can understand easily in this hard-life race. We call the time as "Space Century" and we still wear horse-glasses.

People, who live together, have to share some usual states in daily-life. But they separate from each other in private-life. People have some qualifications that the others have the permission to know about. Although they should share this kind of things in daily-life, they don't need to share their own life with the others.

As I said at the beginning when we miss the important parts, we look for some other events which can fill the lost pieces's place. Maybe, we feel disappointed and try to find some wrong events about the person so that it will carry us to the right way easily. And the best way is to come in to their private-life which contains many mistakes without thinking of our private-life that contain (maybe) much more faults in.

The example- about Bill Clinton- is good but isn't enough to explain my ideas. In my opinion, not only do the president work for country; he also must be a special person whose behaviours could be example for the society. Although I meant "private-life is special and people can do what ever they want", at first, how I can say that I forgot to make an explanation. All famous people don't have to answer all question as "Who does What with Whom?" And really it's not necessary. But except the ones who should carry some important works on their back as a president.

As I told what I thought about this subject now I think I should use an end sentence. I can say that; ignore everything except the ones that will change the world's route one day!



NUMBER:12

PRIVATE -LIFE

Everybody has his own life and this life is very important for one's personality. And so nobody can struggle to learn this private-life

If a person, maybe a president or a taylor has a mistake with his/her life, it is only his/her problem. Everybody is free to choose his life and nobody is responsible from the other's life.

Anyway, private -life must be secret. People can do whatever they want during their lives. Because everyone lives for himself not for the others.

I think most people have mistake in their life like B. Clinton has. But people usually criticise important people's lives. The other's mistakes seem very normal.

And unfortunately in our country singers or actors have same problems. People want them to only do their jobs and want to learn their private life.

For example Bulent Ersoy, the people in our country like talking about her and I think she is very unlucky. She is always criticised by misunderstanding people. I think she can choose whoever she want to marry. He can be young or old. It's her problem

So I think many people can't spend their time as they want. In my opinion it is very bad thing to live for the others

If I'm in another world now, I'll do many things. And I think many people think like I do. let's fly to another world where people don't care each other.

NUMBER:13

I think People need to know more about famous people than their bodies. for Example; Their loves and life habits and exc Because Sometimes. People vote to them (for exaple Clinton) and listen to them. If we know more about them; with a great possible we reach a better agrement. We can andurstand if They say the truth or not. Human`s says onr related with their lifes. If we know their lifes. we easly anderstand their caracter.

Many people seem happy. but in dead some of them have may problems. For instance; about two weeks ago Zerrin Ozer who is a famous pop singer talked to a magazine prograne and said there very significant things. The most importent of these is about sex. If The magazine programe there isn`t nobody knows it.

Yes, and Bill Clinton. The famous sex scandal. I think it`s good for American people to know everything about sex scandal. So that They recorgnise the man. Who decide for them If A metter like this is happen in White House I don`t want ta think anymore. Firstly Bill lied to his family an than the American people. Who knows How many lies there are. Surely it`s good for Americans. After this They will know that only The body on the says aren`t important the most important things is their lifes "life is minnor."

Everywhere we can see if anybody who is famous wants to seam people a good person ofcause wants to publish his an her family photos. Many times we believe this happy scenary. After a few day A magasine programe publish some documents about him. We andurstand what a like man who is.

The famos every time a ster for normal people. many of them want to be like them and live this dream. Because of this the truith news every time must be published buy magasines.



NUMBER:14

In this world every people has a private-life. A president, a student, a butcher, a teacher, a doctor etc. People have different kinds of job. If a teacher has private-life, a president must have private-life. It doesn't mean that a president or a famous singer don't have private life. Some people especially reporters say "Public wants to know famous person's life. So we have to inform public about their life. Their life is very interesting." I don't think like this.

For example Bill Clinton, who is very famous because of his job, everybody knows him. He has private life like everyone. In his job he is very successful president. But he did a wrong relationship with a woman and people blamed him. People want to know what Clinton did with Monica. It's not good. It must not be important for people. It's his private-life. He must find a solution himself for this problem. We can give an example for this topic, Diana. She was a princess of England. She was very famous. But now she died because of people's curiosity. If reporters hadn't behaved her like this, perhaps, she would have lived now. It's not her choice. I think people must be considerateful to the private-life of famous people. Everyday on T.V. people watch a lot of famous people and they are aware of which singer or star are happy, unhappy, married, or unmarried. For example, who loves Tarkan? Where is Tarkan? What does he? And a lot of questions about him. People want to know everything about his/her favourite stars. Their private-life must be unimportant for us. It's his/her life. It's none of our business. People must not put their noise into everything. If so his/her stars can be bored because of these questions. Everybody is free in his/her life. People must live their life how they want.

Therefore; nobody must be care of others' private life. It's people's life they are able to do whatever they want. Everyone must be free.



NUMBER:15

MEANS OF PRIVATE LIFE

People private-life must be secret. Because every human has got different lifestyles. And I think they are free, they can do what they want. We must ignore peoples private-life.

Nowadays Bill Clinton is the best example of this subject. He id something but it doesn't consider us. He did it and he has got some responbolities to his family not the public. American people must look his works not his scandals.

In television we always watch famous artists private life. And it's very boring for watcher. For example she was going to party but she wasn't get his husband with herself. She went there another rich man. Our people can say "we don't want to listen their lifes." Because watching their lifes can't take anything

On the contrary some artists like showing their private-lives. But majority of the famous people try to keep their life-styles.

I think life-style is a means of secret. It can be protect from the society. Every people design a way of life. And they go on that way. On this way nobody can effect the person. Everyone live for themselves. We must think about this.

Bill Clinton made friend another woman. And medie found this material, lateron they added some attractive points. Then most American effect from that event. They change their good-minds about him. They love him because he works for his nation. But that kind of events make a great effect on people. I think it is the best way to hide that things.

At the end I want to say please consider your life not the other people lifes.

Appendix B

Anadolu University, Eskisehir

English for Nonnative Speakers Preparatory Program

Final Exam Grading Standards

5 The essay demonstrates clear competency though there may be a few errors

The content is relevant with the topic and there is no irrelevant content.

The main idea in each paragraph is supported by clear and appropriate evidence/examples.

The paragraphs of the essay are clearly and logically organized.

The text is organized into a clear introduction, body and conclusion.

Few and minor grammar errors.

The use of vocabulary is clear and effective with few inaccuracies.

4 The essay demonstrates competency

Each paragraph has a main idea.

The content is relevant to the topic, but there may be some irrelevant information.

The main idea in each paragraph is supported by appropriate evidence/examples.

The text is organized into introduction, body and conclusion.

The paragraphs of the essay are logically organized.

More grammatical errors in general, a few major errors, which do not interfere with understanding.

The use of vocabulary is clear but not well developed /varied, still with few inaccuracies.

3 The essay demonstrates minimal competency

Some paragraphs do not have a main idea.

The content is generally relevant to the topic, but not all parts of the essay are.

Some main ideas are supported by appropriate evidence/ examples.

Some of the introduction, body and conclusion may be incomplete.

The paragraphs of the essay are not logically organized.

More major errors that sometimes interfere with understanding.

The use of inaccurate vocabulary sometimes confuses the reader.

2 The essay shows a developing competency

Most of the paragraphs do not have a main idea.

Most of the content is not relevant to the topic.

Most of the main ideas are not supported by appropriate evidence /examples.

Some part of the introduction, body and/ or conclusion is missing /incomplete.

The paragraphs of the essay are not organized.

The number and quality of the errors make understanding difficult.

The use of inaccurate vocabulary frequently confuses the reader.

1 The essay demonstrates incompetency in writing

None of the paragraphs has a main idea.

The content is not relevant.

None of the ideas presented are supported by appropriate evidence /examples.

There is no distinct introduction, body and conclusion.

The paragraphs of the essay are not coherent.

Grammatical errors are so frequent that some portions of the essay are incomprehensible.

The use of inaccurate vocabulary makes some portions of the essay incomprehensible.

Appendix C

Instructors' Grades

INSTRUCTOR A

Number of Paper	Content	Organization	Grammar	TOTAL
1	20	10	22	52
2	20	25	18	63
3	15	20	25	60
4	17	15	15	47
5	10	08	15	33
6	10	15	12	37
7	08	10	11	29
8	08	16	30	54
9	25	25	33	83
10	03	05	20	28
11	20	18	33	71
12	25	15	35	75
13	17	15	10	42
14	27	25	37	89
15	20	17	19	66

INSTRUCTOR B

Number of Paper	Content	Organization	Grammar	TOTAL
1	10	15	15	40
2	25	20	25	65
3	20	25	35	80
4	10	15	25	50
5	20	20	25	65
6	25	20	20	65
7	30	25	30	85
8	15	10	25	50
9	25	20	25	70
10	05	05	30	40
11	30	25	35	90
12	30	25	35	90
13	10	15	15	40
14	30	30	35	95
15	30	25	30	85

INSTRUCTOR C

Number of Paper	Content	Organization	Grammar	TOTAL
1	20	15	15	50
2	20	23	15	58
3	27	25	35	87
4	00	15	35	50
5	15	10	30	55
6	20	25	35	80
7	05	15	39	59
8	30	05	32	67
9	20	20	35	75
10	00	05	32	37
11	15	20	38	73
12	25	23	35	83
13	27	20	05	52
14	30	25	38	93
15	28	23	33	84

INSTRUCTOR D

Number of Paper	Content	Organization	Grammar	TOTAL
1	20	20	20	60
2	20	20	15	55
3	15	25	25	65
4	25	25	10	60
5	20	10	15	45
6	20	25	30	75
7	25	25	20	70
8	10	10	20	40
9	20	20	30	70
10	10	15	15	40
11	20	25	20	65
12	15	10	20	45
13	25	20	20	60
14	25	15	20	60
15	20	20	20	60

INSTRUCTOR F

Number of Paper	Content	Organization	Grammar	TOTAL
1	10	10	15	35
2	10	10	10	30
3	20	15	30	65
4	15	10	15	40
5	15	20	15	50
6	20	20	30	70
7	25	25	25	75
8	15	05	10	30
9	20	20	20	60
10	05	05	10	20
11	27	30	35	92
12	25	25	35	85
13	10	05	10	25
14	30	30	35	95
15	20	20	25	65