WRITING PORTFOLIO ASSESSMENT AND INTER-RATER RELIABILITY AT YILDIZ TECHNICAL UNIVERSITY SCHOOL OF FOREIGN LANGUAGES BASIC ENGLISH DEPARTMENT

A Master's Thesis

by

ASUMAN TÜRKKORUR

DEPARTMENT OF TEACHING ENGLISH
AS A FOREIGN LANGUAGE
BILKENT UNIVERSITY
ANKARA

JULY 2005

To my husband Çağrı

for his love, patience and support

&

my parents-in-law for their endless help and support

WRITING PORTFOLIO ASSESSMENT AND INTER-RATER RELIABILITY AT YILDIZ TECHNICAL UNIVERSITY SCHOOL OF FOREIGN LANGUAGES BASIC ENGLISH DEPARTMENT

The Institute of Economics and Social Sciences

of

Bilkent University

by

ASUMAN TÜRKKORUR

In Partial Fulfillment of the Requirements for the
Degree of
MASTER OF ARTS

in

DEPARTMENT OF TEACHING ENGLISH
AS A FOREIGN LANGUAGE
BILKENT UNIVERSITY
ANKARA

JULY 2005

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Teaching English as a Foreign Language.

(Dr. Theodore S. Rodgers)
Supervisor

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Teaching English as a Foreign Language.

(Dr. Susan Johnston)
Examining Committee Member

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Teaching English as a Foreign Language.

(Prof. Dr. Aydan Ersöz)
Examining Committee Member

Approval of the Institute of Economics and Social Sciences

(Prof. Dr. Erdal Erel)
Director

ABSTRACT

WRITING PORTFOLIO ASSESSMENT AND INTER-RATER RELIABILITY AT YILDIZ TECHNICAL UNIVERSITY SCHOOL OF FOREIGN LANGUAGES BASIC ENGLISH DEPARTMENT

Türkkorur, Asuman

MA Department of Teaching English as a Foreign Language

Supervisor: Dr. Theodore S. Rodgers

Co-supervisor: Dr. Susan Johnston

This research study investigated the use of writing portfolios and their assessment by raters. In particular it compared the inter-rater reliability of the portfolio assessment criteria currently in use and the new portfolio assessment criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department. The perspectives of the participants on the portfolio assessment scheme and the criteria were also analyzed. This study was conducted at Yıldız Technical University, School of Foreign Languages, Basic English Department in the spring semester of 2005.

Data were collected through portfolio grading sessions, focus group discussions and individual interviews. The participants in the study were seven

English writing instructors currently working at Yıldız Technical University, School of Foreign Languages, Basic English Department. The instructors scored twelve student portfolios on two different sessions using the criteria customarily used in the institution and the new analytic criteria. Focus group discussions were held before and after the grading sessions. At the end of the grading sessions, instructors were interviewed individually. Grading sessions, focus group discussions and interviews were audiotaped and transcribed.

The inter-rater reliability for both of the criteria types was calculated and found to be marginal. The results of the statistical analysis revealed that there was no difference in results of inter-rater reliability between the groups in both of the grading sessions. However, analysis of the focus group discussion and interviews indicated that instructors would appreciate some form of more standardized, analytic and reliable criteria for portfolio grading.

Key words: Writing portfolio assessment, inter-rater reliability, alternative assessment,

ÖZET

YILDIZ TEKNİK ÜNİVERSİTESİ YABANCI DİLLER YÜKSEK OKULU TEMEL İNGİLİZCE BÖLÜMÜNDE YAZIM PORTFÖYÜ DEĞERLENDİRME SİSTEMİ VE OKUYUCULAR ARASI GÜVENİRLİK

Türkkorur, Asuman

Yüksek Lisans, Yabancı Dil Olarak İngilizce Öğretimi Bölümü

Tez Yöneticisi: Dr Theodore S. Rodgers

Ortak Tez Yöneticisi: Dr Susan Johnston

Temmuz, 2005

Bu çalışma, yazım portföylerinin kullanımını ve onların okuyucular tarafından değerlendirilmesini araştırmıştır. Çalışma özellikle, Yıldız Teknik Üniversitesi Yabancı Diller Yüksek Okulu, Temel İngilizce Bölümü'nde güncel olarak kullanılan portföy değerlendirme kriteri ile, çalışmada önerilen yeni portföy değerlendirme kriterinin okuyucular arası güvenirliğini karşılaştırmıştır. Katılımcıların portföy değerlendirme sistemi ile kriterler üzerine görüşleri de analiz edilmiştir. Çalışma, 2005 bahar yarıyılında, Yıldız Teknik Üniversitesi Yabancı Diller Yüksek Okulu, Temel İngilizce Bölümü'nde yürütülmüştür.

Veriler portföy değerlendirme oturumları, odak grup tartışmaları ve bireysel görüşmeler aracılığıyla toplanmıştır. Araştırmaya Yıldız Teknik Üniversitesi, Yabancı Diller Yüksek Okulu, Temel İngilizce Bölümünde çalışmakta olan ve yazım dersleri veren yedi öğretim görevlisi katılmıştır. Öğretim görevlileri iki farklı portföy değerlendirme oturumunda, hem kurumda kullanılmakta olan kriteri hem de yeni analitik kriteri kullanarak on iki öğrenci portföyü değerlendirmişlerdir. Değerlendirme oturumlarının öncesinde ve sonrasında odak grup tartışmaları gerçekleşmiştir. Değerlendirme oturumlarının sonunda ise bireysel görüşmeler yer almıştır. Değerlendirme oturumları, odak grup tartışmaları ve görüşmeler teybe kaydedilmiş ve yazıya dökülmüştür.

Her iki kriter türüne ait okuyucular arası güvenirlik, iki değerlendirme oturumundan elde edilen notlar kullanılarak hesaplanmıştır. İstatistiki analiz sonuçları iki değerlendirme otumunun okuyucular arası güvenirlik sonuçlarında herhangi bir fark olmadığını göstermiştir. Ancak, odak grup tartışmaları ve görüşmeler incelendiğinde, öğretim görevlierinin portföy değerlendirmesinde bir çeşit standart, analitik ve güvenilir kritere sıcak baktıkları görülmüştür.

Anahtar kelimeler: Yazım portföyü değerlendirmesi, okuyucular arası güvenirlik, alternatif değerlendirme,

ACKNOWLEDGEMENTS

I would like to express my gratitude for my thesis advisor, Dr Theodore S. Rodgers, for his on-going guidance and contribution to this study. I am deeply grateful to him for his endless support and understanding at times of trouble, and for his confronting attitude towards all the potential problems.

I owe much to Dr Susan Johnston, the director of the MA TEFL program, for being with me even at the most intimate times. Without her assistance, understanding and her big heart full of love for others it would have been impossible to complete the program.

I am thankful to committee members Dr. Susan Johnston and Dr. Aydan Ersöz who enabled me to benefit from their expertise and to make necessary additions to my study.

I would like to thank Dr William Snyder for sharing his deep knowledge and experience at the early stages of formulating the thesis topic.

I would like to take this opportunity to thank all the members of MA TEFL faculty, Dr Susan Johnston, Dr Theodore S. Rodgers, Michael Johnston, Dr Ian Richardson, Prof Dr Engin Sezer and Dr Ayşe Yumuk Şengül, for sharing their profound knowledge through the courses they have given.

I wish to thank all my classmates in the MA TEFL Program with whom I had the greatest enjoyment of sharing and learning together. I would especially like to thank my dearest friend, Pınar Uzunçakmak for her encouraging and soothing words whenever I needed to hear them, even at the toughest times.

I am thankful to the former director of Yıldız Technical University School of Foreign Languages, Ins. Perihan Akbulut, for the encouragement and the permission to attend the MA TEFL program. I would also like to thank the former head of department, Ins. Serap S. Alıcı, for her support when I asked for a leave to attend the MA TEFL program. I would like to thank, the head of the Basic English Department, Ins. Aylin Alkaç, for her support and encouragement to conduct this study. I appreciate especially instructors Cemile Güler and Hande Abbasoğlu's help in the implementation of the study. I would also like to thank instructors Ebru Demirtaş, Narin İlkkılıç, Meliha Kesemen, Özlem Mendi Donduran, Özlem Sezen, Tanju Sarı and Zeynep Çalım for their precious participation in the study.

I am grateful to my family and my husband's family. Without their love, constant encouragement, kindness, help and affection my life could not have been that easy during the program.

Finally, I am deeply grateful to my husband, Çağrı, for his love, patience and encouragement that always made me feel strong, even from miles away. He is the light of my life.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

CHAPTER I: INTRODUCTION

Introduction

Student assessment has a number of forms, including traditional tests and alternative assessment types. Systematic alternative assessment forms, such as portfolios, peer assessment, and self-assessment have been used in language learning contexts since the mid-eighties.

Portfolio assessment, as an alternative assessment option, has been used to evaluate both oral and written communication and discourse (Wiig, 2000). The portfolio is a purposeful, integrated collection of student work that shows student effort, progress, or achievement in a given area over time (Paulson, Paulson & Meyer, 1991; Genesee & Upshur, 1996). It includes a wide variety of work samples, such as writing samples, book reports, film reviews, short stories, students' samples of recorded speech, written self-evaluation, journals, teacher's notes and reports, and other pieces of work of the students' own choice (Georglou and Paulov, 2002). In a portfolio process, students develop self-reflection and self-monitoring and they become actively involved in their own language learning process by helping to set the focus, establish the standards, select contents, and judge merit of student products (Paulson & Paulson, 1994).

Writing is an essential skill in academic language contexts because writing contributes to the development of higher cognitive functions such as analysis and synthesis, which is also the principal way in which students report what they have learned. Writing instruction and assessment have undergone considerable changes over the last thirty-five years (Raimes, 1991). According to Dinçman (2002), writing instruction was formerly based on grammar drills, worksheets and sentence diagramming as ways to improve composition in the classroom. However, there have been changes in the approach to writing. These changes in approach include process writing, journal reflections, projects, timed writing, whole language instruction, and portfolios.

The use of portfolio assessment for writing in the English as a foreign language (EFL) context has grown rapidly at educational institutions during the last twenty years (Gussie & Wright, 1999). For example, pilot studies of the European Language Portfolio (ELP) models have made it possible for member states of Council of Europe to implement portfolios in different academic areas. Since Turkey is hoping to become part of the European Union, ELP implementations have been launched at various high schools and at universities (Oğuz, 2003). According to a survey of the pilot ELP scheme in Turkey, participating teachers and students have indicated strong positive responses. The teachers agreed that the ELP makes a positive contribution to the language teaching and learning process and develops learner motivation and autonomy (Demirel, 2004).

Reliability, which is a critical issue in any kind of assessment method, relates to the consistency of the results of an assessment method (Bachman & Palmer 1996; Hamp-Lyons, 1996). Reliability in portfolio assessment by instructors seeks a standardization of criteria, particularly in any large-scale assessment process (Song & August, 2002). To ensure quality-grading procedures, the implementation process for portfolios should be carefully designed from the beginning to the end, with the criteria matching the institutional goals and objectives. Further, instructors need to be informed about, take part in and be trained about the evaluation process and assessment scales (Lumley & McNamara 1993; Hamp-Lyons, 1996). The decision-making process, and content and the assessment criteria should provide reliability and fairness in marking for all students across classes.

In the assessment of writing across a program, inter-rater reliability is a significant issue. Inter-rater reliability is the degree of similarity of assessment marks given by different reader-raters (Henning, 1993). Inter-rater reliability can be promoted by having two or more raters evaluate the same writing sample and then compare their marks and criteria (Hyland, 2003). Since portfolio assessment is a relatively new procedure, the reliability issue must be seriously taken into consideration.

Because of this increased importance put on portfolios in evaluation schemes, the purpose of this research is to find the inter-rater reliability of the criteria that are currently being used and the criteria to be proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department. In addition, instructors' perspectives on the portfolio assessment implementation, their own personal criteria, and the proposed analytic criteria will be explored.

Background of the Study

Alternative assessment, authentic assessment, and performance assessments are labels for proposals to provide options to traditional assessment methods by further promoting student creativity and performance on significant tasks (Ewing, 1998). According to Brown and Hudson (1998), traditional assessment types are selected-response assessments consisting of test items like true-false, matching, and multiple choice questions, and constructed-response assessments including fill in, short answer test items and timed performance assessments. Alternative assessments are personal-response assessments including essays, writing samples, diaries, oral discourse, exhibitions (Ewing, 1998); portfolios, conferences, self-assessments, and peer assessments (Brown & Hudson, 1998).

Alternative assessments are said to enhance student creativity and productivity, provide qualitative data about both the strengths and weaknesses of students, encourage open disclosure of standards and rating criteria, promote the use of meaningful instructional tasks, and call upon teachers to perform new instructional and assessment roles (Brown & Hudson, 1998). The focus on process as well as product and dedication to a longitudinal assessment approach are the main determinants of the decisions that educators make in implementing alternative assessments. This is especially so in writing classes.

Therefore, the use of portfolio assessment is increasing, particularly in the assessment of writing. Hamp-Lyons and Condon (1993) assert that portfolio-based assessment is superior to traditional assessment because of the many programmatic benefits it brings with it.

Portfolios in language learning are also an important issue as stated in the ELP. ELP presents a format that makes it possible for students to document their progress in multi-lingual competence by recording learning experiences of all kinds over a range of languages. ELP is a personal type of portfolio aiming to motivate learners by helping them realize their efforts to expand language skills at all levels and to provide a record of the linguistic and cultural skills they have acquired. In terms of pedagogy, ELP functions to enhance the motivation of the learners, to help learners plan their learning and reflect on their own learning process (Schneider & Lenz, 2001). The ELP takes into account the diversity of learner needs according to age, learning purposes, contexts, and background. The basic division of ELP is in three parts: *The Language Passport* provides "an overview of the individual's proficiency in different languages at a given point in time" (Schneider & Lenz, 2001, p.16). *The Language Biography* facilitates the "learner's involvement in planning, reflecting upon and assessing his or her learning process and progress" (Schneider & Lenz, 2001, p. 19). *The Dossier* offers "the learner the opportunity to select materials to document and illustrate achievements or experiences recorded in the Language Biography or Language Passport" (Schneider & Lenz, 2001, p. 38).

Apart from the individual ELP portfolio described above, most other portfolios are institution-based. Establishing a portfolio-based writing assessment necessitates careful planning and continuous checking. In her study, Nunes (2004) focuses on two basic principles in developing portfolios. The first principle is that a portfolio should be dialogic and facilitate on-going interaction between teacher and students. It should include teacher feedback and revised, edited and rewritten forms of student writing samples. The second principle is that portfolios should document the reflective thought of the student. Through reflective thinking in writing, students can develop a more responsive relationship with their own learning process. Therefore portfolios should not only be considered as a source of examples of student work to be assessed but as a "self-contained learning environment with valid outcomes of its own" (Paulson & Paulson, 1994, pg. 15).

Reading, evaluating and scoring portfolios constitute the most important steps towards achieving reliability in portfolio evaluation. As Hamp-Lyons and Condon (1993) emphasize, portfolio assessment requires "as much of an evaluative stance and attention as a traditional essay-test does" (p. 187). This requirement necessitates the need for assessment criteria. In order for a program to be fully accountable for its decisions, it must have explicable, sharable and consistent criteria (Hamp-Lyons & Condon, 1993). According to Brown and Hudson (1998), credibility, auditability, multiple tasks, rater training, clear criteria, and triangulation of any decision-making procedures along with varied sources of data are important ways of improving the reliability and validity of assessment procedures used in any educational institution.

Portfolios allow a more detailed look at a complex activity because they contain several samples collected over time and texts written under different conditions. They are therefore generally considered to be more valid than traditional assessment methods (Hamp-Lyons, 1991). Reliability in portfolio assessment involves ensuring reliability across raters, promoting objectivity, preventing mechanical errors that would affect decisions and standardizing the grading process (Brown & Hudson, 1998). As in Brown and Rodgers' (2002) model, using more than one experienced rater to carry out the assessment independently can enhance inter-rater reliability.

Statement of the Problem

Portfolios are becoming more widely used in English language programs in Turkish universities as an alternative assessment method to traditional tests. However, as this qualitative approach to student assessment becomes more common, it is necessary to determine if the actual assessment of the portfolios by instructors is reliable.

As in all other forms of assessment, the designers and users of alternative assessment must make every effort to structure the ways they design, pilot, analyze, and revise the procedures so that the reliability and validity of the procedures can be studied, demonstrated, and improved (Brown & Hudson, 1998). Developing clearly and well-designed writing portfolio assessment criteria can help to encourage objectivity in instructors to approach a higher reliability in their analysis of student writing. If instructors assess writing samples without making use of such criteria, the assessment system lacks a basic element which should be addressed by the program administration.

The writing program at Yıldız Technical University, School of Foreign Languages, Basic English Department has been implementing portfolio assessment for three years. Every year there is obvious development in the practice of this alternative assessment in terms of portfolio design, portfolio contents and teacher feedback techniques. Besides portfolios, student writing is also assessed through four achievement tests, one mid-term examination and a final writing exam. In these exams students are required to write an essay, a letter or a story in a given time. Evaluation rubrics are prepared for each examination according to the genre of the writing piece. During the academic year writing instructors have to read hundreds of papers; therefore, teachers who do not teach writing are required to score examination papers, except for the final writing exam. The final writing exam is scored by two experienced raters who also are writing instructors.

Although instructors at Yıldız Technical University, School of Foreign Languages, Basic English Department use a trial rating scale for assessing writing exam papers, there is no criteria for the assessment of writing portfolios. Because of this lack of standardized criteria, there might be significant differences between the scores given by two instructors on the same portfolio. In order to improve the quality of the writing program, the administration asked the researcher to conduct a research study on the reliability of writing portfolio assessment. Therefore, this study aims to determine if there are significant differences between scores given by different instructors on the same portfolio. The study will also identify the inter-rater reliability for an alternative portfolio-based assessment scale proposed for Yıldız

Technical University, School of Foreign Languages, Basic English Department. Instructors' perspectives on portfolio assessment implementation in the institution and on the use of both of the scales will also be examined.

Research Questions

1. What is the inter-rater reliability of Basic English teachers using the "traditional" writing portfolio assessment criteria prescribed at Yıldız Technical University, School of Foreign Languages, Basic English Department?
2. What is the inter-rater reliability of Basic English teachers using the new writing portfolio assessment criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department?
3. What are the instructors' general perceptions of the writing portfolio scheme at Yıldız Technical University, School of Foreign Languages, Basic English Department?
4. What are the instructors' perceptions of the use of the "traditional" writing portfolio assessment criteria presently used at Yıldız Technical University, School of Foreign Languages, Basic English Department?
5. What are the instructors' perceptions of the use of writing portfolio assessment criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department?

Significance of the Problem

Students are asked to prepare portfolios in their writing courses at Yıldız Technical University, School of Foreign Languages, Basic English Department. Since portfolios have a 5% value in the overall student grade and play an important role in their graduation from the preparatory program, reliable writing portfolio assessment criteria are needed.

The use of standardized and reliable criteria will encourage objectivity in instructors and fairness for the students. Inconsistencies between the rater scores may be reduced.

By presenting an alternative writing portfolio assessment scale and the results of an inter-rater reliability study on instructors' evaluations using the new writing portfolio assessment criteria at Yıldız Technical University, School of Foreign Languages, Basic English Department, this study might be useful for EFL instructors, curriculum designers and program administrators who are implementing portfolio assessment in their institutions. The results of the study may help them to identify the problems that affect the reliability of the assessment and to develop assessment measures that are appropriate to portfolio design and reliable across instructor-raters.

Conclusion

In this chapter, an overview of the literature on writing portfolio assessment and inter-rater reliability has been provided. The statement of the problem, research questions, and the significance of the study have also been presented. In the second chapter relevant literature is explored. In the third chapter the methodology of this research study is presented. In the fourth chapter, the analysis of the data is given. In the last chapter, conclusions are drawn from the data in the light of literature.

CHAPTER II: LITERATURE REVIEW

Introduction

This research study investigates the use of writing portfolios and their assessment by raters. In particular, it seeks to compare the inter-rater reliability of the portfolio assessment criteria currently in use and the new portfolio assessment criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department. The study partially focuses on the assessment of student writers on the basis of portfolios, which contain samples of student writing, collected throughout the term. There is a major section examining the literature on various aspects of portfolio assignments and assessment. Incidental discussion on the portfolio issue appear in sections throughout this survey, where these appear most naturally to fit.

This chapter reviews the literature relevant to portfolio assessment. The chapter consists of four sections. First, the concept of assessment of language performance will be reviewed. Second, issues on writing in the second language (L2) classroom will be presented. This section will be followed by a section on reliability theory in assessment and factors involving inter-rater reliability. The last section covers portfolios, including information about their history, types, pros and cons as instructional instruments and their use in assessment.

Assessment of Language Performance

Assessment and evaluation play a critical role in students' educational progress. Evaluation is considered the broader term, assessment being considered a form of evaluation. Language evaluation not only encompasses learner proficiency, but also represents a critique of the language program, materials and teaching effectiveness (Council of Europe, 2001).

Language learning is a creative activity whereby learners process and produce oral and written discourse based on the rules of a language system which they have internalized (Hendrickson, 1984). Assessment of language performance, in other words *performance assessment*, requires the learner to create written or oral language products or performances (Council of Europe, 2001).

Since it is difficult to measure what mental processes students undergo while producing spoken or written language, evaluation tools need to be carefully designed in acknowledgement of the inaccessibility of mental operations (Breland, 1996). Brown (1986) points out that evaluation models should be qualitative, context-rich, and naturalistic. The aim of evaluation tools should be to understand specific cases, rather than general truths, and involve multiple sources of information about students' strengths and weaknesses (Brown, 1986).

Gronlund (1998) asserts that a carefully designed assessment program can help language learning in various ways. First, assessment can influence student motivation by providing them with clear goals and tasks to be mastered and by giving feedback about language progress. Second, assessments can promote student "self-assessment" since they provide models and criteria of learning progress. This information about student progress helps provide insights into their language abilities. Assessments also provide feedback about educational efficacy in terms of the realization of instructional goals, the methods and materials used, and the learning experiences of the learners.

Types of assessment can be grouped under two broad headings: standardized assessment and alternative assessment. These types will be explained in detail below.

Standardized Assessment

According to Brown and Hudson (1998) standardized assessments or traditional assessments are selected response assessments including test items such as true-false, matching and multiple choice questions, and constructed response assessments include fill-in, short answer questions and some traditional tasks like essay writing.

In *Standardized Assessment Primer* by Association of American Publishers (www.publishers.org) it is stated that the purpose of standardized tests is to provide valid and reliable information to educators, students, parents and policymakers. For educators and the public, standardized tests provide information that helps them work on the following issues (p. 4):

1. Identify the instructional needs of individual students so educators can respond with effective, targeted teaching and appropriate instructional materials;
2. Respond with effective, targeted teaching and appropriate instructional materials;
3. Judge students' proficiency in essential basic skills and challenging standards and measure their educational growth over time;

4. Evaluate the effectiveness of educational programs;
5. Monitor schools for educational accountability.

According to Gottlieb (2000), traditional, standardized, and norm-referenced assessment has never been an especially reliable or valid indicator of L2 learners' knowledge or ability. However, Henning (1991) states that many performance assessment programs that obtain high levels of rater reliability are, in fact, standardized assessments, based on examinees' performing the same tasks under the same conditions. In such assessments, raters can be trained with benchmark sample performances of the identical tasks used in the assessment instrument. As I will suggest, this has not often been the case in non-standardized or alternative assessment types, such as portfolio assessments.

In terms of writing courses, standardized testing assesses students by means of a limited range of writing samples—or no writing samples at all—which may give insufficient or misleading information about student's actual ability. According to Tierney et al. (1991), standardized tests in writing are also disadvantageous in other ways. Scoring may be largely mechanical and often performed by inexperienced or untrained raters. Standardized assessment focuses on product rather than process and necessarily assesses all students on the same dimensions. Moreover, standardized assessments do not allow opportunities for writer revision, which indicates that the writer may or may not be capable of learning from his or her errors.

Alternative Assessment

All language tests are forms of assessment, but there are also many forms of performance assessments, such as checklists, used in continuous assessment or informal teacher observations, which are not described as tests (Council of Europe, 2001). Such forms of assessment comprise a somewhat loose category variously labeled as alternative assessment, authentic assessment or performance assessment. Discussions of these "alternatives" have dominated the testing literature since the 90s (Ewing, 1998). In this discussion, "alternative assessment" is contrasted to traditional, standardized assessment.

The term alternative assessment is often used as an "umbrella" term for any "non-traditional" assessment (Brindley, 2001; Butler, 1997, p. 5). Alternative assessments have produced several assessment approaches called "performance assessment," "alternative assessment," and "authentic assessment." Tedick and Klee (1998) state that these assessment types are different from traditional assessments both in structure and scoring; learners are expected to perform meaningful tasks showing what they can do, and learning is viewed as a process with performance evaluated according to specific criteria. Herman et al. (1992, as cited in Butler, 1997) summarize these multiple definitions:

> We use these terms (alternative assessment, authentic assessment, and performance-based assessment) synonymously to mean variants of performance assessments that require students to generate rather than choose a response. Performance assessment by any name requires students to actively accomplish complex and significant tasks, while bringing to bear prior knowledge, recent learning, and relevant skills to solve realistic or authentic problems. Exhibitions, investigations, demonstrations, written or

> oral responses, journals, and portfolios are examples of the assessment alternatives we think of when we use the term "alternative assessment."(p. 5)

A critical rationale behind alternative assessment is the belief that not all learners learn in the same way, and "learning does not occur in a straight line" (Butler, 1997, p. 4). One source of assessment information about learner proficiency is not enough and maybe unreliable; thus, each learner should be assessed in multiple ways so that he or she can demonstrate their language abilities in different forms. The second basis of alternative assessment is that feedback comes not only from teachers, but also from peers or the students themselves in order to enhance learning (Butler, 1997).

Alternative assessment has been seen as appropriate in assessing skills of reading, writing, speaking, researching, problem solving, and original invention. Leeming (1997) lists some of the important tenets of alternative assessments:

- Assessment should examine the processes as well as the products of learning.
- Assessment should promote higher-level thinking and problem solving skills.
- Assessment should integrate assessment methodologies with instructional outcomes and curriculum content
- Specific criteria and standards for judging student performance should be set.
- An integrated and active view of learning requires the assessment of holistic and complex performance.
- Assessment systems that provide the most comprehensive feedback on student growth include multiple measures taken over time (p.51).

Different alternative assessments vary in the scoring and interpretation of the assessments. Using checklists and rubrics for assessing student performance on various language tasks is one primary form of alternative assessment (Tedick & Klee, 1998). Checklists are used to observe student performance and work over time. They are also used to determine whether a specific criterion is present. Rubrics, on the other hand, focus on the quality of written or oral performance. Rubrics are created on the basis of four different scale types (Tedick & Klee, 1998): holistic, analytic, primary-trait, and multi-trait which were originally developed for large scale writing assessment. These scales will be discussed in more detail in the 'Assessment of L2 Writing' section.

Encouraging reflection through self-assessment and peer assessment is another aspect of alternative assessment. Students need to self-assess in order gain understanding of their own learning. Barnhardt et al. (1998) state that in the portfolio process, student self-assessment promotes critical thinking and responsibility in students. Students are able to grade themselves depending on their weaknesses and strengths. Self-assessment also allows teachers to see how students view their progress leading to instruction that is individualized in response to specific student needs (Barnhardt et al., 1998).

Peer assessment is used when students evaluate each other's work depending on pre-determined objectives and rating scales. Using peer-assessment in the portfolio process promotes "cooperation, trust, and a sense of responsibility, not just to oneself but to others" (Barnhardt et al., 1998, p. 63). It is recommended that peer-assessment in the portfolio process should include at least two student pairs (Tedick & Klee, 1998).

Portfolio assessment, as will be discussed in the final section of this chapter, ideally encompasses all that has been discussed above: it emphasizes a variety of tasks that elicit spontaneous as well as planned language performance for a variety of purposes and audiences. There is a use of rubrics to assess performance, and a strong emphasis on self-reflection and self-assessment and peer assessment (Tierney et al., 1991, Tedick & Klee, 1998).

As mentioned before, tasks used in any kind of alternative assessment should give students the opportunity to show what they can do with the language. Alternative assessments are criterion-referenced assessments, and the type of task varies according to the language skill. To exemplify alternative assessment methods, it is possible to include videos of role-plays (Butler, 1997; Tedick & Klee, 1998); interviews, group or individual presentations; debates and information-gap activities in speaking and listening tasks; journals, compositions, letters, e-mail correspondence or discussions; skimming authentic tasks for gist, scanning for specific information, analyzing articles or stories by different authors, for different audiences in reading tasks (Tedick & Klee, 1998); research reports, experiments, portfolios in writing (Ewing, 1998).

Criticisms about alternative assessments focus on three main issues: *validity* (whether an assessment tests what it aims to test), *reliability* (whether the results of an assessment would be the same when applied to the same examinees over time), and *objectivity* (whether an assessment is free from biases) (Butler, 1997). Other challenges alternative assessments face are the adaptation processes of teachers and students and providing the appropriate learning and assessment environment (Tedick & Klee, 1998). Both teachers and students who are used to traditional assessment types need to be informed and trained about alternative assessment types in these processes. They may react in a negative or uninformed way to their new roles. Students will need training on how to reflect on their own performance as well as how to give useful feedback to their peers' performance. A cooperative learning environment needs to be created because students need to reflect on their own learning process and give feedback to their peers in a comfortable, relaxed, constructive atmosphere. Thus, alternative assessments should be carefully designed and implemented (Tedick & Klee, 1998).

Cole et al. (2000) point out that educators believe that assessment should measure student performance in relation to educational goals which have been previously agreed to by the student and evaluator. Alternative assessment builds a strong bridge between learning and evaluation and, in fact, is often closely integrated with instruction (Douglas, 2000).

Butler (1997) emphasizes that implementing alternative assessment requires a change in the curriculum, too. Learning is not viewed as filling learners with an amount of information, but as a process in which learners are involved actively in their own development and in which teachers assume roles as facilitators rather than bankers of information. This approach is said to lead to more "learner-centered pedagogy", which supports collaboration between teacher and student in terms of power and responsibility in the educational process (Tedick & Klee, 1998, p.2). Students then become more active in their own learning process. While students are involved in the learning and evaluation processes, teachers become developers of learner-centered activities. This implementation results in alternative assessment methods which allow students to be more closely involved in the evaluation process and to reflect on their own learning as a result of this involvement (Tedick & Klee, 1998).

Writing in the L2 Classroom

Writing is a complex activity in which the writer demonstrates a range of knowledge and skills. This complexity makes it unlikely that the same individual will perform equally well on all occasions and on all tasks (Hyland, 2003). Writing effectively is not purely a matter of choosing vocabulary and mastering grammar and memorizing rhetorical forms. It is a process that requires writers to gather ideas, provide coherence between ideas, have an argument, and address a prospective reader's questions, objections or expectations (Leeming, 1997). Because of this complexity it has been argued that an appropriate way of assessing L2 writing should be found which more accurately reflects this complexity. It is within this ongoing discussion, that proposals moving away from traditional standardized testing towards alternative assessment types have been forwarded.

Research in L2 writing has focused on 3 main dimensions: “a) features of the texts that people produce; b) the composing processes that people use while they write, c) the socio-cultural contexts in which people write” (Cumming, 2001, p. 3).

In terms of text features, research supports the view that as second language learners’ proficiency increases, the complexity and accuracy of sentences and vocabulary improve, and learners become more competent in organizing their ideas according to appropriate genre forms (Cumming, 2001). Research on the composing processes suggests that as people learn to write in a second language, they are better able to plan, revise, and edit their texts effectively. In respect to the influence of socio-cultural contexts in L2 writing, Cumming (2001, p. 8) observes “L2 writers are required to write in various contexts such as universities, colleges, community settings, working environments. They become aware of the ways of cooperating with people from different discourse communities”.

Types of L2 Writing

A list of types of writing is almost without limit, including labels, lists, letters, reminder notes, bulletin board announcements, banners, songs, editorials, novels and declarations. Attempts to classify writing types vary from a traditional, primary school inventory of narrative, expository and persuasive writing styles to sophisticated analysis of academic genres (Swales, 1990). One classification that has found some favor with those teaching second language learners was that proposed by Roman Jacobson and adapted by Rodgers (1989, as cited in Brown & Rodgers, 2002, pp.40-42). In this categorization the various genres are grouped by the language function that the genres typically serve. An abbreviated form of this classification with writing examples is shown below:

1. *Emotive function* focuses on the feelings of the message *sender*.
   Genres: Valentines, graffiti, confessions
2. *Referential function* focuses on the message *content.*
   Genres: Textbooks, news broadcasts, encyclopedias, recipes
3. *Metalinguistic function* focuses on the linguistic *code.*
   Genres: Grammars, dictionaries, thesauri
4. *Poetic function* focuses on artistry of message *composition*.
   Genres: Novels, songs, poems
5. *Phatic function* focuses on the social *contact*.
   Genres: Social notes, birthday cards, invitations
6. *Persuasive function* focuses on influencing the *receiver*.
   Genres: Advertisements, sermons, infomercials

Probably all of these types of writing appear as practice exercises in various handbooks on the teaching of L2 writing. From the perspective of this study, many of these appear as possible writing types comprising a writing portfolio which may or may not be analyzed and graded. I will return to a consideration of writing types in the section on portfolios.

Assessment of L2 Writing

Hyland (2003) argues that assessment is not simply administering exams and giving scores. Moreover, evaluating students' writing performance is a formative process which has a strong impact on student learning, the writing course design, teaching strategies and teacher feedback. Writing assessment tools vary in type, ranging from class tests, short essays, long project reports, and writing portfolios to large-scale standardized examinations.

There are four principal types of scoring scales for rating essays—holistic, analytic, primary trait and multi-trait. Holistic scoring evaluates the language performance as a whole (Cohen, 1994). Each score on a holistic scale represents an overall impression of the potential language abilities (Tedick & Klee, 1998). A true holistic reading of an essay involves reading for an individual impression of the quality of the writing, by comparison with all other writing the reader sees on that occasion (Hamp-Lyons, 1996). This approach generally focuses on what is done well. However, Cohen (1994) lists a number disadvantages associated with holistic scales. Firstly, one single score is not considered suitable to interpret students' strengths and weaknesses. Secondly, holistic scoring is a sorting or ranking procedure and is not designed to offer correction, feedback, or diagnosis for learners. Scores generated in this way cannot be explained easily, either to the other readers who belong to the same assessment group and who are expected to score reliably together, or to the people affected by the decisions made through the holistic scoring process (Hamp-Lyons, 1991). Third, the scores may cause a misinterpretation of students' sub-skills. It is also difficult for raters to give equal weighting to all aspects in each paper and to produce fair results. A sample holistic scoring is given below in Figure 1.

Figure 1

Holistic Scale for Assessing Writing

4	Excellent—Communicative; reflects awareness of sociolinguistic aspects; well-organized and coherent; contains a range of grammatical structures with minor errors that do not impede comprehension; good vocabulary range.
3	Good—Comprehensible; some awareness of sociolinguistic aspects; adequate organization and coherence; adequate use of grammatical structures with some major errors that do not impede comprehension; limited vocabulary range.
2	Fair—Somewhat comprehensible; little awareness of sociolinguistic aspects; some problems with organization and coherence; reflects basic use of grammatical structures with very limited range and major errors that at times impede comprehension; basic vocabulary used.
1	Poor—Barely comprehensible; no awareness of sociolinguistic aspects; lacks organization and coherence; basic use of grammatical structures with many minor and major errors that often impede comprehension; basic to poor vocabulary range.

(Tedick & Klee, 1998, p. 31)

Analytic scoring requires the use of separate scales, each assessing a different feature of writing (Cohen, 1994). Each subcategory is scored separately and scores are then added up for an overall score (Tedick & Klee, 1998). Analytic scoring is advantageous in that it prevents raters from collapsing the sub-categories during scoring and provides a useful tool for rater training (Cohen, 1994). However, there is a possibility that the raters will not use each part of analytic scale properly since rating on one scale may influence rating on another (Cohen, 1994). Additionally, research finds little evidence that "writing quality is the result of the accumulation of a series of sub-skills" (Cohen, 1994, p. 319). Below in Figure 2 is an analytic ESL composition scoring profile by Jacobs et al. (1981, as cited in Hughes, 2003, p. 104), which is also used as the proposed analytic criteria in this study.

Figure 2

Analytic Scoring Scale

Content	
30-27	Excellent to very good: knowledgeable - substantive - thorough development of the thesis - relevant to assigned topic
26-22	*Good to* average: some knowledge of subject – adequate range - limited development of thesis - mostly relevant to topic, but mostly lacks detail
21-17	Fair to poor: limited knowledge of subject - little substance - inadequate development of topic
16-13	Very poor: does not show knowledge of subject - non-substantive - not pertinent - OR not enough to evaluate
Organization	
20-18	Excellent to very good: fluent expression - ideas clearly stated/supported - well-organized - logical sequencing - cohesive
17-14	Good to average: somewhat choppy - loosely organized but main ideas stand out - limited support - logical but incomplete sequencing
13-10	Fair to poor: non-fluent - ideas confused or disconnected - lacks logical sequencing and development
9-7	Very poor: does not communicate - no organization - OR not enough to evaluate
Vocabulary	
20-18	Excellent to very good: sophisticated range - effective word/idiom choice and usage - word from mastery - appropriate register
17-14	Good to average: adequate range - occasional errors of word/idiom form, choice, usage, but meaning not obscured
13-10	Fair to poor: limited range - frequent errors of word/idiom form, choice, usage - meaning confused or obscured
9-7	Very poor: essentially translation - little knowledge of English vocabulary, idioms, word form - OR not enough to evaluate
Language Use	
25-22	Excellent to very good: effective complex constructions - few errors of agreement, tense, number word order/function, articles, pronouns, prepositions
21-18	Good to average: effective but simple constructions - minor problems in complex constructions - several errors of

	agreement, tense, number, word order/function, articles, pronouns, prepositions but meaning seldom obscured
17-11	Fair to poor: major problems in simple/complex constructions - frequent errors of negation, agreement, tense, number, word, order/function, articles, pronouns, prepositions and/or fragments - meaning confused or obscure
10-5	Very poor: virtually no master of sentence construction rules - dominated by errors, does not communicate, OR not enough to evaluate
Mechanics	
5	Excellent to very good: demonstrates mastery of conventions - few errors of spelling, punctuation, capitalization, paragraphing
4	Good to average: occasional errors of spelling, punctuation, capitalization, paragraphing but meaning not obscured
3	Fair to poor: frequent errors of spelling, punctuation, capitalization, paragraphing - poor handwriting - meaning confused or obscured
2	Very poor: no mastery of conventions - dominated by errors of spelling, punctuation, capitalization, paragraphing – handwriting, OR not enough to evaluate

Primary trait rubrics are based on a view that one can only judge whether a writing sample is good or not by reference to its exact context, and that appropriate scoring criteria should be developed for each prompt (Hamp-Lyons, 1991). The primary trait approach gives detailed attention to specific aspects of writing and it allows focus on one issue at a time; however it could be difficult for raters to focus exclusively on one specific trait in scoring (Cohen, 1994). Another disadvantage of the primary trait approach is that a specific aspect of writing may not deserve to be considered "primary" (Cohen, 1994). A sample primary trait rubric is given below.

Figure 3

<u>Primary Trait Rating Scale</u>

Primary Trait: Persuading an Audience	
0	Fails to persuade the audience.
1	Attempts to persuade but does not provide sufficient support.
2	Presents a somewhat persuasive argument but without consistent development and support.
3	Develops a persuasive argument that is well developed and supported.

(Tedick & Klee, 1998, p. 35)

Finally, in multi-trait scorings, the rater considers a number of aspects of the essay, but not in the same way they do in analytic scoring (Cohen, 1994; Grabe & Kaplan, 1996). In this approach the traits represent “specific aspects of writing of local importance” and validity is improved because “the test is based on expectations in a particular setting” (Cohen, 1994, p. 323). It is believed that this approach has a positive impact on teaching and learning. However, it is a challenge for the trait developers to identify and validate traits that are appropriate for each given context (Cohen, 1994). A sample multi-trait rubric is given below.

Figure 4

Multi-trait Rubric

	Main Idea/Opinion	*Rhetorical Features*	*Language Control*
5	The main idea in each of the two articles is stated very clearly, and there is clear statement of change of opinion.	A well-balanced and unified essay, with excellent use of transitions.	Excellent language control, grammatical structures and vocabulary are well chosen.
4	The main idea in each article is fairly clear, and change of opinion is evident.	Moderately well balanced and unified essay, relatively good use of transitions.	Good language control; and reads relatively well, structures and vocabulary generally well chosen.
3	The main idea in each of the articles and a change of opinion are indicated but not so clearly.	Not so well balanced or unified essay, somewhat inadequate use of transitions.	Acceptable language control but lacks fluidity, structures and vocabulary express ideas but are limited.
2	The main idea in each article and/or change of opinion is hard to identify in the essay or is lacking.	Lack of balance and unity in essay, poor use of transitions	Rather weak language control, readers aware of limited choice of language structures and vocabulary.
1	The main idea of each article and change of opinion are lacking from the essay.	Total lack of balance and unity in essay, very poor use of transitions.	Little language control, readers are seriously distracted by language errors and restricted choice of forms.

(Cohen, 1994, p.330)

Song and August (2002) assert that the writing abilities of English as a Second Language (ESL) students are more difficult to assess than those of native speakers. ESL students' writing is more appropriately evaluated in large-scale assessments like portfolios. Hamp-Lyons and Condon (2000) also support the idea that portfolios are suitable for ESL students since they supply a broader view of students' writing abilities and provide a better alternative to timed exams. According to research results, it has been found that students from different cultural and educational backgrounds brought different expectations and strategies to the timed writing exams and responded in different ways with different levels of success (Hamp-Lyons and Condon, 2000).

Writing samples of students are assessed by two main approaches: direct and indirect assessment (Grabe & Kaplan, 1996; Hyland, 2003). Largely due to problems caused by reliability issues in direct assessment of L2 writing assignments, various indirect assessment methods have been proposed. Indirect assessment tools such as multiple-choice questions or cloze tests allow the students to demonstrate grammar and sentence construction skills, which are elements in successful writing. Indirect assessment forms have been used in large-scale standardized examinations like TOEFL and are often preferred because they are considered to allow standardization, reliability and flexibility in administration and scoring (Hyland, 2003). On the other hand, direct assessment, which is based on the production of written texts, is considered to be more valid and authentic. Direct writing assessments are subjective measurements of written essays. The direct approach can evaluate both composition and basic skills. It is believed that direct writing assessment has face validity, but requires subjective measurement often resulting in rater disagreement (Schwarz & Collins, 1995).

Recently, in writing skill assessment there has arisen an approach of using free-response writing tasks, in contrast to traditional standardized assessment. This approach has had a broad impact on writing skill assessment (Breland, 1996). Many United States (US) based national examinations and testing programs, such as the Graduate Management Admission Test (GMAT), the Graduate Record Examination (GRE), National Assessment of Educational Progress (NAEP) and the Medical College Admission Test (MCAT), have added free-response essay assessments. However, some testing programs like the Scholastic Assessment Test (SAT), the Test of General Education Development (GED) and Writing Skills Test (WST) have not followed this practice or are doing so only in moderation (Breland, 1996, p. 2).

Reliability

According to Bachman and Palmer (1996), the most important feature of a test is its 'usefulness'. They define usefulness as "… a function of several different qualities, all of which contribute in unique but interrelated ways to the overall usefulness of a given test" (p. 18). These different qualities are reliability, construct validity, authenticity, interactiveness, impact, and practicality. Test developers need to find an appropriate balance among these qualities according to their purpose, students, and situations (Karslı, 2002).

Barnhardt et al. (1998) define reliability as the consistency and accuracy of the assessment tool to measure students' performance. According to Henning (1991) reliability refers to the capacity of the assessment procedures to guide raters to rank-order the same samples of writing performance consistently in the same way. Hyland (2003) defines a writing assessment task as reliable as long as it measures consistently the same student on different occasions and the same task across different raters.

There are many factors apart from the test itself that cause variations in student scores. Some factors might be the physical conditions of the exam room, time of day, the rubric and instructions, and the prompt genre (Hyland, 2003). Gronlund (1998) adds that a limited number of items in tests and a limited range of scores also lower the reliability of test scores.

Henning (1991) lists possible causes for low reliability of scoring. First, several aspects of the scoring systems may contribute to the lack of reliability. Unclear or inconsistent terminology in the scoring rubrics could contribute to error in scoring. Insufficient training may also contribute to low reliability. Finally, the nature of alternative assessments—in particular, the lack of standardization of tasks and administrative conditions—may undermine reliability.

In terms of performance assessment Gronlund (1998) lists the factors that lower the reliability as follows. “Insufficient number of tasks, poorly structured assessment procedures, inadequate scoring guides and scoring judgments that are influenced by personal bias” (p. 219) are those that affect the reliability of scoring in performance assessments. In order to avoid these factors, a sufficient number of samples should be taken; assessment procedures should define the nature of tasks, the assessment conditions and the criteria; the candidate’s choice of topics and genres should be restricted; appropriate scoring rubrics that describe the criteria should be used; and judges need to be trained (Gronlund, 1998; Hughes 2003).

Barnhardt et al. (1998, p. 28) state that reliability can also be supported through “triangulation” which requires data about a specific language skill from different sources. Considering this quality, portfolios are accurate tools since they provide feedback about the learner’s progress from the learner, peers and the teachers.

Lumley and McNamara (1993) relate reliability issues in test scoring especially to rater factors. They note that differences between idealized raters and actual raters are regrettable but unavoidable. Differences between judges could be understood in terms of overall severity or randomness in rating consistency. Harper and Misra (1976, as cited in Lumley & McNamara, 1993) found that, of these two elements, the extent of random error was as great as the extent of differences between the mean scores allocated by a panel of judges and more problematic since it is harder to anticipate and eliminate.

Reliability in portfolio assessment involves establishing clear and detailed criteria for both the portfolio and the contents of the portfolio before students undertake their assignments (Barnhardt et al., 1998). Other ways to promote reliability in portfolios involve ensuring reliability across raters, promoting objectivity, preventing mechanical errors that would affect decisions and standardizing the grading process (Brown & Hudson, 1998).

Types of Reliability

Brown & Rodgers, (2002) discuss two types of reliability: They claim that *person-related reliability* should ensure that the person is prepared and understands what is expected, and *instrument-related reliability* can be achieved by using different methods of assessment and insuring optimal assessment conditions.

Hyland (2003) states that reliability in scoring student writing has two considerations. 1. *Inter-rater reliability*, which requires that all raters agree on the scoring of same student performance. This type will be discussed in the next section in more detail. 2. *Intra-rater reliability* is provided when the same rater scores the same student performance in the same way on different occasions. Intra-rater reliability is the consistency of the judgments by the same rater on two occasions. Brown (1996) argues that raters' remembering their scores from the first administration can confound the results of reliability estimates. As a result of this possible problem, this form of reliability is not as often discussed in language testing as inter-rater reliability.

Inter-rater Reliability

Research supports that writing raters are influenced by many factors and can weight the writing subcategories differently during the scoring of student papers (Hyland, 2003). One rater focuses on content and communicative clarity, whereas the other uses grammatical accuracy as the sole criterion for rating (Bachman, 1990). One might be influenced by the handwriting or page length while the others look for organization.

Using more than one experienced rater to carry out portfolio assessment independently can enhance assessment reliability (Barnhardt et al., 1998). In order to reduce rater variability, Lumley and McNamara (1993) suggest implementing rater-training sessions in which raters are introduced to the assessment criteria and asked to rate a series of selected performances. During these sessions, ratings are carried out independently and raters become aware of the extent to which they rate similarly or dissimilarly with other raters and try to achieve a common interpretation of the rating criteria. The training session is followed by additional follow up ratings and the reliability of the scores is again analyzed. Only after these training sessions, should raters and rating panels be selected. It has been found that rater training can reduce the extent of rater variability in terms of overall severity and random errors and can help develop self-consistency in raters (Lumley & McNamara, 1993).

Hamp-Lyons (1996) asserts that training rater-readers is not an easy issue. In order to provide valid and reliable scorings of writing there are various aspects to take into consideration: "The context in which the training occurs, the type of training given, the extent to which training is monitored, the extent to which reading is monitored, and the feedback given to readers" (p. 82).

Reliability of Teachers as Writing Evaluators

Assessing student papers is one of the most important responsibilities of writing teachers because the decisions they make about how they give grades affect students' lives, as do other forms of student evaluation. Williams (1998) defines three of the most important topics in writing assessment by teachers as being: validity, reliability, and time. Validity is related to matching what one is teaching to the assessments students are asked to take part in. Both teaching and writing are complex and multi-faceted. Finding valid matches between instruction and assessment is difficult even for assessment professionals. Reliability is related to the consistency of evaluation. If an assessment procedure is reliable, then the evaluation process will not be affected by any outside factors, such as the evaluator or the time and place of administration. Time is of central importance to teachers who are already heavily burdened. A feasible assessment procedure should not occupy a great deal of a teacher's time.

A study by Anderson, Bachor and Baer (2001) reveals that the evaluation of student achievement is not an easy process. Their study involved 127 pre-service elementary school teachers who assessed the performance of three "simulated" students on 6 language arts tasks. The portfolio structure was developed so that each portfolio contained the work of the three simulated different students on six language arts tasks. Each student teacher was required to mark each of the six products of the three students and then submit a final mark and lettergrade for each student; however, they were not provided with criteria, keys or rubrics. They were also required to keep a journal and record their thoughts they had about scoring the portfolios. The analysis of the data shows that final marks are not the same thing as final lettergrades although they are closely related. Individual teachers sometimes use additional information in creating letter grades that is not necessarily reflected in numerical final marks. The results also indicated the potential for the portfolio approach to collecting information about the evaluation of student achievement by teachers.

Hamp-Lyons (1996) states that different readers respond to different facets of writing. Research findings support that readers respond to cultural differences in essays, or rater behavior can vary according to sex, race or geographic origin. These variations have led an emphasis on rater training in writing assessment programs (Hamp-Lyons, 1996).

Often the only evaluators of students' writings are teachers. Hyland (2003) states that teachers need assurance that they are scoring student performance ethically and reliably. They also expect to see that there is consistency between their scores and those that other teachers might give to the same writing performance (Hyland, 2003). Hughes (2003) emphasizes that the scoring of student writings should not be allocated to inexperienced raters. Therefore he suggests that the scores after each administration be analyzed and raters whose scores result in inconsistency not be used again.

Portfolios

Portfolios are collections of multiple samples of student writing, written and collected over time and represent students' abilities and learning progress. Bushman & Schnitker (1995) state that portfolios are concerned with the process of learning and student's language awareness as well as products of learning. Portfolios encourage language awareness since they include reflection and self-evaluation of student work.

Portfolios enable students to display their writing abilities in a more natural and less stressful way. Portfolios represent multiple samples of student writing abilities and may include drafts, reflections, readings, diaries, observations of genre use, teacher or peer responses, as well as finished texts (Hyland, 2003).

Portfolios should not only be considered as sources of examples of student work to be assessed (Herman et al., 1993). If correctly implemented, students may become increasingly independent learners as a result of the portfolio process, and the outcomes may be more valid in reflecting their own interests (Paulson & Paulson, 1994). Thus, portfolio reading and response requires "as much of an evaluative stance and attention as a traditional essay-test does" (Hamp-Lyons & Condon, 1993, p. 187). This requirement creates the need for democratically achieved and widely agreed upon assessment criteria. In order for an educational program to be fully accountable, it must have explicable, sharable, consistent criteria (Hamp-Lyons & Condon, 1993). These include credibility, auditability, multiple tasks, rater training, clear criteria, and triangulation of any decision-making procedures along with varied sources of data. Brown and Hudson (1998) emphasize important ways to improve the reliability and validity of the assessment procedures used in educational institutions.

Portfolio Contents

There are two major types of portfolio models; one being portfolios that include every work the student has produced, the other being portfolios that include only selected samples of student work. These samples may be student or teacher designated in accordance with the course objectives. Portfolios can represent language performances in different genres with or without drafts revisions and finished products (Hyland, 2003).

As previously noted, a single writing performance cannot fairly reflect or measure a skill as complex as writing ability (Daiker et al., 1996). Therefore, the portfolio contents typically include multiple samples of writing from a number of occasions, a variety of kinds or genres of writing, and students reflections on their portfolios, writing processes, and on themselves as writers (Daiker et al., 1996).

Brown (2004) gives a detailed list of some materials included in portfolios:

- several drafts and final forms of essays and compositions
- reports, project outlines,
- poetry and creative prose,
- artwork, photos, newspaper or magazine clippings,
- audio and/or video recordings of presentations, demonstrations,
- journals, diaries and other professional reflections,
- tests, test scores, and written homework exercises,
- notes on lectures, and
- self- and peer-assessments—comments, evaluations and checklists (p. 256).

As well, portfolios may contain copies of writing assignments, students' responses to each other, reflection papers and final summative essays (Douglas, 2000). Some works in portfolios can be assigned, others may be self-initiated; some are long-term projects, some are one page writings (Santos, 1997).

Background of Portfolio Assessment

Assessment of student progress in school has been an important part of education affecting students, parents, teachers, administrators and even educational policy makers. Students are administered tests and other assessment tools to monitor their progress and to provide feedback. At this point it is important to point out how portfolios became a tool of assessment in education.

In 1993 the United States Department of Education created a call for a shift from “mastery of minimum competencies to promotion of excellence in education” (Gussie & Wright, 1999, p. 4), the National Council on Education Standards and Testing recommended the development of an assessment system in order to:

> exemplify for students, parents, and teachers the kinds and levels of achievement that should be expected; improve classroom instruction and the learning outcomes of all students; inform students, parents and teachers about students’ progress towards the national goals, measure and hold students, schools, districts, states, and the nation accountable for educational performance (Gussie & Wright, 1999, p. 5).

This attempt to change the existing system led to a new assessment system focusing on student development of meta-cognitive skills (such as critical thinking, self-monitoring, and self-assessment) as well as student ability to execute a rich variety of performance tasks. Instead of relying on single source of information about student strengths and weaknesses, educators moved towards alternatives in assessment. Portfolio assessment, particularly, received major attention. As a result, today portfolios are used in many academic areas including mathematics, chemistry, physics, teacher training and English for academic purposes (Douglas, 2000).

Advantages of Portfolio Assessment

Portfolios offer a number of benefits for both teachers and students. Portfolios can be considered as a powerful assessment approach since they are said to re-shape the roles of teachers, students and the assessment process in a positive way.

According to Brown and Hudson (1998), portfolio assessment strengthens student learning by increasing learners' attention and involvement in their learning processes and promoting student-teacher and student-student collaboration. A portfolio of student work can help students develop ownership of their learning and can encourage self-analysis as they reflect on their work (Wortham, 1998; Trotman, 2004; Grabe & Kaplan, 1996; Genesee & Upshur, 1996). Students often have the freedom and responsibility to select the content of the portfolio and the conditions for their writing which also promotes motivation, learner autonomy and critical thinking.

Shober (1996) investigated how a portfolio can be used to present growth in students' narrative writing and how portfolios can be used as a discussion tool for parent/teacher/student conferences. The study was conducted in a twelve-week period with 22 students from the fourth grade target group. The students completed three writing samples during this period which were assessed for growth and understanding of the writing process. During the completion of the three writing samples planning, prewriting, drafting, conferring and revising writing processes were actively practiced. Sharing the portfolio with the parents and teacher/parent/student conference was a major part of the study Evaluation conferences were held between teacher and student, student and a peer, or in a small writing group. Results of the study indicated that 68% of the students showed improvement in narrative writing. However, although the questionnaire results demonstrated the positive attitudes of parents towards portfolios and conferences, only 55% parental participation in the conferences was achieved.

Portfolio assessment can change the teacher's role from that of an error-hunter and challenger to that of a guide. By using portfolio assessment, teachers tend to focus more on process rather than product, facilitating students as they engage in planning, drafting, feedback, collaboration and revision. As Brown and Hudson (1998) indicate, portfolios provide unique insights into the progress of each student. Portfolios also help teachers and program designers plan further instruction and learning experiences for the students since they provide detailed data for integration of student learning (Wortham, 1998). Furthermore, portfolio assessment allows an integration of curriculum and assessment; that is, there is the possibility for a continuous, developmental and fair evaluation in relation the program goals as well as documentation for rethinking these goals.

Grabe and Kaplan (1996) emphasize that the "process" movement in writing instruction was linked to portfolio considerations. They point out that one of the major positive impacts of the writing process approach has been the rethinking of responses to student writing. Student revision and teacher response has become central at all stages of the writing process: pre-writing, first drafting, revising, and final-draft writing.

Portfolio assessment is said to improve assessment processes by enhancing student and teacher involvement in assessment processes and allowing the assessment of multiple dimensions of language learning (Brown & Hudson, 1998; Genesee & Upshur, 1996). It allows for assessment of multiple writing samples across a range of topics and task types (Wortham, 1998). Portfolios require students to perform significant tasks and directly demonstrate competence by constructing, rather than selecting responses (Ewing, 1998). Portfolio assessment typically provides samples of the best work that a student is capable of producing.

Validity, an important quality of any assessment approach, is another strength of portfolio assessment. Teachers are able to make inferences from judging a collection rather than judging a single piece of work (Trotman, 2004). Furthermore, portfolios are directly related to what is taught and what students are able to do in response to instruction. Because they contain several samples and because they can be constructed so that texts written under different conditions are included, portfolios allow a more complex look at a complex activity, and are therefore generally considered to be more valid (Hamp-Lyons, 1991).

Portfolio assessment reflects program goals and therefore also provides feedback to program administrators on how clear these goals are and to what extent they are being achieved. These various characteristics make portfolio assessment a potentially strong assessment tool. The portfolio is not simply a collection of a student's work, but a meaningful measure of student progress. It has been stated that no other assessment approach promotes reforms in the teacher's role, student learning and the assessment process as effectively as portfolios (Brown & Hudson, 1998).

Challenges of Portfolio Assessment

We have seen that the use of portfolio assessment in writing not only avoids over-reliance on student performance in a single timed exam, but also promotes writing instruction and assessment with validity, authenticity, interactivity and washback (Trotman, 2004). However, there are disadvantages of using portfolio assessment as well. Brown and Hudson (1998) address five issues that can challenge portfolio implementation: design decisions, logistics, interpretation, reliability and validity.

Design decision issues deal with the instructor deciding the content and grading criteria. The questions “who will decide upon the content” and “who will specify the purposes” are challenging questions that institutions face (Brown & Hudson, 1998; Trotman, 2004). Institutions have to decide how much they will allow instructors and students to direct the decision making process. Establishment of grading criteria is also a crucial issue since it has been found that (Hamp-Lyons & Condon, 1993) portfolio readers often lack explicit criteria and standards to measure portfolios.

Overcoming logistical issues, such as time constraints, is another main concern in implementing portfolio assessment. Portfolio assessment is time-consuming and increases the workload of teachers (Oğuz, 2003). A research study by Bushman and Schnitker (1995) points out that time management was the biggest obstacle in implementing portfolios. Teachers are engaged in helping students in their planning, editing and revising stages. Continuous interaction between teacher and students during the portfolio development process requires teachers to spend more time and dedication to supporting this process.

Setting the standards in grading, providing fairness to each student, and training teachers to make fair evaluations are interpretation issues that challenge portfolio assessment. Portfolio grading should reflect student achievement and success as represented in their portfolios (Brown & Hudson, 1998). As Gearhart and Herman (1998) state in their study on large-scale portfolio assessment, a portfolio rater should be familiar with the student and the classroom context to score a student's portfolio collection. Research by Webb (1993, as cited in Gearhart & Herman, 1998) suggests that an individual's performance as part of a group activity may or may not reflect his or her true capability. A rater's score for a portfolio may overestimate student performance because it constitutes a rating of efforts that were teacher or student assisted. The study indicates that low-ability students had higher scores on the basis of group work than on individual work. Thus, training teachers on the implementation, assessment and interpretation of portfolios represent confounding concerns for institutions implementing portfolio assessment.

Another drawback in the grading of portfolios deals with reliability of portfolio assessment. Variation in inter-rater scoring is the most common issue affecting reliability. Inter-rater reliability involves "determining the correlation of two or more raters for the same writing samples, and then adjusting the obtained coefficients" (Henning, 1991, p. 286). By such adjustments, inter-rater reliability can be improved. In most institutions, due to time constraints, one teacher marker evaluates the writing portfolios with the assumption that the teacher is familiar with the students and the tasks. Only in rare circumstances are teacher raters trained for fair and objective judgment and are two raters employed to score the portfolios. If an assessment system is not reliable, it cannot be valid.

Validity, on the other hand, includes determining how adequately the portfolios exemplify student work, development, and abilities. Another critical validity issue is whether the contents of the portfolios are appropriate to the goals of the course. Perfectly acceptable writing samples may not be congruent with instructional objectives. Herman, Aschbacher, and Winters (1992, as cited in Ewing, 1998, p. 11) emphasize that "quality assessment should meet certain common standards and they offer the criteria developed by the Center for Research, Evaluation, Standards, and Student Testing" as worthy standards for increasing validity and reliability particularly as these apply to portfolio assessment.

Portfolio Assessment

Portfolio assessment is a performative assessment. It is becoming a more common type of assessment in writing programs since it allows students to demonstrate development of their writing products over time. Portfolios also act as a process-oriented assessment of long-term progress in writing since they provide evidence of editing and revision in the construction of a final product (Douglas, 2000). Therefore, portfolio assessment is seen as both product and process assessment (Hirvela & Pierson, 2000).

As previously indicated, portfolio assessment is one type of alternative assessment. Among alternative assessment types there has been continuing growth of interest in and practice of portfolio-based assessment of writing. Hamp-Lyons and

Condon (1993) assert that portfolio-based assessment is superior to traditional holistic assessment because of the many programmatic benefits it brings with it. As Brown and Hudson (1998) note, portfolio assessments enhance student creativity and productivity, provide information about both the strengths and weaknesses of students, encourage open disclosure of standards and rating criteria, use meaningful instructional tasks, and call upon teachers to perform new instructional and assessment roles.

Hyland (2003) lists the procedures for designing and implementing portfolio assessment, the first being the determination of the content of portfolios based on course objectives and student needs analysis. Second, it is crucial to discuss the purposes and procedures of the portfolios with students regularly throughout the term. A discussion and decision on the assessment criteria among the teachers will be helpful and should be shared with the students. Planning the draft check dates and feedback conferences is a further step that helps keep students on task in the portfolio productions. Writing products and presentations enable students to share their works with the others. Finally teachers need to encourage reflection on the part of students so that they can analyze their own writing and even reflect on the criteria decided for portfolios.

Douglas (2000, p.243) suggests five characteristics of portfolio assessment procedures:

1. Comprehensive: both depth and breadth of work is represented
2. Predetermined and systematic: careful planning is essential

3. Informative: work must be meaningful to teachers, students, staff and parents
4. Tailored: work included must relate to the purpose of assessment
5. Authentic: work should reflect authentic contexts, in and out of the classroom

Criteria for Assessing Portfolios

Specific scoring criteria need to be carefully discussed among teachers and writing program administrators. Criteria used in the assessment of portfolios should particularly strive to demonstrate language development (Douglas, 2000).

Gronlund (1998) emphasizes that the criteria should define the type of performance to be assessed and the intended learning outcomes to be achieved. The standards also define the levels of acceptable performance. In terms of portfolios, the criteria will include not only text features, but also dimensions of thinking and self-reflection, and perhaps, others (Hamp-Lyons & Condon, 2000). The standards and the criteria are then used in preparing rating scales or scoring rubrics to evaluate the portfolio work samples.

The first step of establishing the portfolio criteria should be consultation between the administration and the faculty of the institution (Larson, 1996). Groups of educators can discuss and compare standards for the criteria (Murphy & Grant, 1996). As Hamp-Lyons and Condon (2000) emphasize, specific scoring criteria need to be carefully negotiated. Since it is difficult for teachers to leave their own criteria aside and get used to the new criteria. Gronlund (1998) states that students should be informed about the criteria and standards by which their performance will be evaluated. Moreover, Gronlund suggests that students get involved in the decision process of setting the criteria and the preparation of the rating scales.

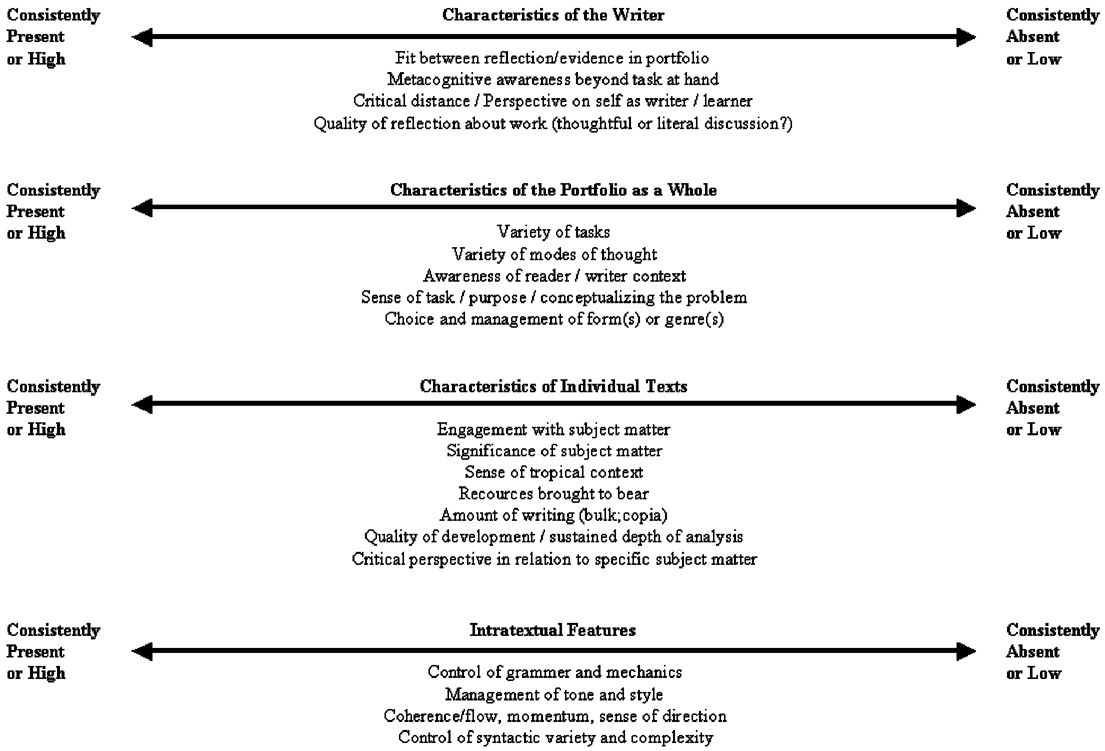
Scoring of portfolios is sometimes problematical. There are two main approaches to grading portfolios: holistic and multi-trait. As discussed before holistic scoring is the most common form of scoring for large scale or in-class writing assessments and is achieved by reading a text and deciding on a general, subjective score (Grabe & Kaplan, 1996).

Multi-trait scoring is believed to have many advantages in portfolio-based assessment, (Hamp-Lyons & Condon, 2000) and is a more common and preferred option than single trait scoring for writing assessment. Hamp-Lyons (1991) suggests that the traits can reflect different types of texts, stages of the revised drafts, purposes of writing and more.

The holistic method may be effective with smaller samples, but it is unlikely to be reliable with longer and more open portfolios which display considerable variation. The multi-trait option more faithfully reflects the complexities of both the products and the processes involved, but may become unwieldy if too many different criteria are scored. Hamp-Lyons and Condon (2000) suggest a useful heuristic for devising criteria based on main elements to be assessed.

Figure 5

Dimensions for assessing portfolios



Consistently Present or High | Characteristics of the Writer | Consistently Absent or Low

Fit between reflection/evidence in portfolio
Metacognitive awareness beyond task at hand
Critical distance / Perspective on self as writer / learner
Quality of reflection about work (thoughtful or literal discussion?)

Consistently Present or High | Characteristics of the Portfolio as a Whole | Consistently Absent or Low

Variety of tasks
Variety of modes of thought
Awareness of reader / writer context
Sense of task / purpose / conceptualizing the problem
Choice and management of form(s) or genre(s)

Consistently Present or High | Characteristics of Individual Texts | Consistently Absent or Low

Engagement with subject matter
Significance of subject matter
Sense of tropical context
Recources brought to bear
Amount of writing (bulk;copia)
Quality of development / sustained depth of analysis
Critical perspective in relation to specific subject matter

Consistently Present or High | Intratextual Features | Consistently Absent or Low

Control of grammer and mechanics
Management of tone and style
Coherence/flow, momentum, sense of direction
Control of syntactic variety and complexity

(Hamp-Lyons & Condon, 2000, p. 144)

It is very important to develop clear criteria for the overall quality of the portfolio. These criteria should be shared, discussed and understood by the students before finalizing their portfolio (Santos, 1997).

Conclusion

This chapter reviewed the literature on assessment of language performance, writing in the L2 classroom, reliability, and portfolios. The next chapter will focus on methodology, which covers participants, instruments, procedures and data analysis used in the study.

CHAPTER III: METHODOLOGY

Introduction

The purpose of this study is to investigate portfolios as an alternative assessment system to assess writing in the L2 classroom, as well as inter-rater reliability of teachers as writing evaluators. The study is conducted at Yıldız Technical University, School of Foreign Languages, Basic English Department. The answers to the following research questions are given in the study:

1. What is the inter-rater reliability of Basic English teachers using the "traditional" writing portfolio assessment criteria prescribed at Yıldız Technical University, School of Foreign Languages, Basic English Department?
2. What is the inter-rater reliability of Basic English teachers using the new writing portfolio assessment criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department?
3. What are the instructors' general perceptions of the writing portfolio scheme at Yıldız Technical University, School of Foreign Languages, Basic English Department?
4. What are the instructors' perceptions of the use of the "traditional" writing portfolio assessment criteria presently used at Yıldız

Technical University, School of Foreign Languages, Basic English Department?

5. What are the instructors’ perceptions of the use of writing portfolio assessment criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department?

This chapter outlines the methodology selected for this study and gives information about the participants, instruments, data collection procedures, and data analysis.

Participants

The participants involved in this research study are seven writing instructors working at Yıldız Technical University, School of Foreign Languages, Basic English Department. There are 120 instructors currently working at Yıldız Technical University, School of Foreign Languages, Basic English Department 21 of who teach writing. In order to obtain a representative group of teachers, one third of this population was selected for this detailed study on the basis of willingness and experience in teaching writing.

Thus, teachers who were available for the initial study were asked if they would participate in the study; they agreed. Seven of the writing instructors volunteered to participate and signed the consent form (see Appendix A). The background information about the participants is presented in Table 1 as follows:

Table 1

The participants of the actual study

Total years teaching experience	Less than 1 year	1-3	4-6	7-10	Above 10
Number of teachers			4	2	1

Teaching experience at YTU	Less than 1 year	1-3	4-6	7-10	Above 10
Number of teachers		1	3	3	

Teaching experience in writing at YTU	Less than 1 year	1-3	4-6	7-10	Above 10
Number of teachers	1	4	2		

In the actual study, six of the participants are female and one of the participants is male. The participants' years of experience in teaching English ranged from four to more that 10 years. Their years of experience in teaching writing ranged from one to six years. All participants have experienced implementing portfolio assessment in writing for at least 1 year.

Instruments

In order to look at the inter-rater reliability of two writing portfolio assessment criteria the following instruments are used: student portfolios, the writing portfolio assessment criteria currently used at Yıldız Technical University, School of Foreign Languages, Basic English Department, the analytic criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department, audio recordings of focus group discussions and individual interviews and scores given by each participant to each student portfolio.

Student Portfolios

For the actual study, portfolios of 12 students were selected by the researcher. Six of the portfolios were scored in the first grading session and the other six were scored in the second grading session by the raters. Portfolios were selected from different classes than the participant instructors' in order to avoid subjectivity of judgment. They represented the upper, middle and lower range of student work. All portfolios were completed in the first term of 2004-2005 academic year.

The content of the portfolios were reduced from seven to five items due to the scoring time constraints. The items consisted of first and final drafts of four compositions and one letter. In the actual study only the final drafts were scored. The types of texts were as follows:

'Daily routine of the writer or a famous person'
'Description of a house'
'A letter to a friend from holiday'
'Writing a story based on picture cues'
'Good and bad sides of a favorite sports'

The Writing Portfolio Assessment Criteria Used at Yıldız Technical University, School of Foreign Languages, Basic English Department.

The writing portfolio assessment criteria currently being used at Yıldız Technical University, School of Foreign Languages, Basic English Department is actually an unwritten, traditional one. The writing instructors are expected to score the first and second drafts of each item in the portfolios and give an overall portfolio grade according to their subjective criteria.

The Analytic Criteria Proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department.

The scoring profile for ESL compositions by Jacobs et al. (1981, as cited in Hughes, 2003, p. 104) was used in this study as the alternative criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department (see Appendix B). Analytic criteria for writing portfolio assessment were adopted for three main reasons. First, analytic criteria allow the scoring of different sub-skills, thus the irregular development of sub-skills in individuals can be graded accordingly. Secondly, scorers are required to consider aspects of performance that they might otherwise ignore. Thirdly, the scorer has to give a number of scores for each category and this will tend to make the scoring more reliable (Hughes, 1989, as cited in Karslı, 2002).

Audio Recordings of Focus Group Discussions

Another instrument used in this study was audio recording of focus group discussions. The instructors held three focus group discussions in Turkish. The first focus group discussion was held on the first day of the portfolio grading sessions. It was a twenty-minute discussion as a “warm-up session” to portfolios, their contents and importance. The researcher asked instructors questions on issues such as the implementation of portfolio assessment, the assigned 5% value of portfolios in the overall grades of students, the effects of portfolio assessment on students’ writing abilities and performance, and questions on portfolio contents and their suggestions on various contents that could be included in the future.

The second focus group discussion was held after the first grading session. In the thirty-minute discussion instructors were asked questions about what goes into the grading in portfolio assessment, the criteria they use in grading portfolios, their weights on different sub-components of their criteria, their perspectives on their own portfolio assessment criteria and problems they had that affected the reliability of the assessment.

The final focus group discussion was held after the second and "analytic" grading session. In this thirty-minute focus group discussion the participants talked about their perceptions of the new criteria. The extent of the difference in their grading decisions between the first and the second day were discussed. Follow-up questions were asked in accordance following the direction of the discussions and interviews. The audio recordings of the focus group discussions were transcribed, coded and necessary segments were translated into English.

Audio Recordings of Individual Interviews

Finally, the participants were interviewed individually at the end of the second grading session. The interviews were used in order to get information about the participants' perceptions and attitudes on portfolio implementation, views on teachers as writing evaluators and on the two grading methods they used in the grading sessions. The researcher encouraged L1 use in the interviews in order for the teachers to express themselves more unreservedly. The audio recordings of individual interviews were transcribed, coded and necessary segments were translated into English.

In the interviews, the researcher asked six questions (see Appendix C). The interview questions focused on portfolio assessment implementation in their institution, the need for adequate training on portfolio assessment in writing classrooms, the consistency between teacher evaluations of student portfolios and the comparison of the two scales used in both grading sessions.

Scores Given by Each Participant to Each Student Portfolio

In order to look at the inter-rater reliability of the subjective criteria and the alternative analytic criteria, each participant's scores to each of the 12 student portfolios were analyzed. Statistical analysis was used to compute inter-rater reliability in the two portfolio grading systems.

Data Collection Procedures

In January I requested and received permission from Yıldız Technical University, School of Foreign Languages, Basic English Department to conduct my research.

The pilot study was done on the 1st of March with three MA TEFL students. I piloted the discussion and interview questions in order to ensure that all of the questions were clear, focused on the topic and of the right length. I also asked for my pilot teachers' suggestions, however, they stated that there was no need to make any changes or additions.

The seven participant instructors experienced two writing portfolio-grading sessions on two different days, the first one on the 8th and the second one on the 10th of March. In order not to be affected, the participants were not told the focus of the study. The researcher only explained the general process of the research study to the participants.

At the beginning of the first grading session the participants had the first focus group discussion on writing portfolio assessment. Afterwards, the instructors scored six portfolios with their subjective criteria in the way they had always done in one hour-thirty five-minutes. Of the five items included they graded only the final drafts. After the grading session the second focus group discussion was held. Teachers were asked to weight the “assumed” sub-components of their subjective writing assessment criteria. The sub-components were content, organization, vocabulary, language use and mechanics. Raters assigned a percentage of “importance” weight to each of these sub-components and discussed how these weights influenced their scores in the fist grading session. The grading session and the two focus group discussions were completed in a two and a half-hour period.

On the second day of the portfolio grading sessions, the researcher presented the analytic criteria by Jacobs et al. (1981) and had the instructors discuss and agree on the sub-component weights. Discussing the importance of each sub-component weight, the participants decided the weights of the sub-components together. After this agreement was reached, the instructors graded the other six portfolios according to the new analytic criteria in a two-hour period. Afterwards, in the third focus group discussion the comparison of the two scales and instructors’ perspectives on both of the portfolio assessment criteria were discussed. This second session was completed in three and a half hours. Individual interviews took place immediately following this assessment session in order to help the participants recall the details of the grading sessions without difficulty.

Data Analysis

The data analysis was completed in two stages. First, the scores given to six student portfolios using the subjective assessment criteria by the seven participants and the scores given to the other six student portfolios using the alternative analytic criteria by the seven participants were used to calculate inter-rater reliability.

Second, the focus group discussions and interviews were analyzed and coded by focusing on the participants' perceptions of writing portfolio assessment the criteria used in both portfolio grading sessions, and the problems that the raters faced that would affect the reliability of assessment.

The data analysis procedures and results will be explained in more detail in the following chapter.

Conclusion

This chapter on methodology gives general information about the aim of the study, listing the research questions the researcher attempted to answer. It also provides information about the participants of the study, instruments used, data collection procedures, and data analysis. In the next chapter, I present the data analysis done using the above-mentioned qualitative and quantitative methods to answer the research questions.

CHAPTER 4: DATA ANALYSIS

Introduction

This study investigates writing portfolios as an alternative assessment system to assess writing in the L2 classroom. As well, it examines inter-rater reliability of teachers as writing evaluators of writing portfolios at Yıldız Technical University, School of Foreign Languages, Basic English Department. The study also investigates instructors' opinions about writing portfolio assessment in their institution. The collected data were analyzed to answer the following research questions.

1. What is the inter-rater reliability of Basic English teachers using the "traditional" writing portfolio assessment criteria prescribed at Yıldız Technical University, School of Foreign Languages, Basic English Department?
2. What is the inter-rater reliability of Basic English teachers using the new writing portfolio assessment criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department?
3. What are the instructors' perspectives of the writing portfolio scheme at Yıldız Technical University, School of Foreign Languages, Basic English Department?

4. What are the instructors' perspectives on the use of writing portfolio assessment criteria presently used at Yıldız Technical University, School of Foreign Languages, Basic English Department?
5. What are the instructors' perspectives on the use of writing portfolio assessment criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department?

The results of the analysis will be presented in three main sections. In the first section the analysis of teachers' scores on portfolios in the two grading sessions is presented. In the second section the results of the focus group discussions are analyzed in order to explore the instructors' perceptions of the portfolio assessment and the assessment criteria used in both of the sessions. In the third section results of individual interviews are discussed.

Analysis of Instructors' Scores

In order to investigate the inter-rater reliability of the subjective criteria currently used and the analytic criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department, two sets of data were collected through portfolio grading sessions. In the first grading session, seven writing-instructor raters scored six writing portfolios using the subjective "traditional" criteria prescribed by the department. After discussion and re-design of the criteria, the same seven raters scored another six portfolios using the "new" analytic criteria in the second grading session.

Research Question 1: Inter-rater Reliability of the Subjective Criteria

After the first grading session the scores of the instructors given to six writing portfolios, using their subjective criteria were collected. The scores are shown in Table 2.

Table 2
Portfolio grading with subjective criteria

Portfolio Sample	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Range	Mean	SD
1	60	75	65	70	80	92	100	40	77.42	14.42
2	65	65	55	60	70	78	90	35	69.00	11.77
3	70	75	70	70	75	75	85	15	74.28	5.34
4	75	76	70	70	90	98	100	30	82.71	12.99
5	65	70	60	70	65	70	90	30	70.00	9.57
6	65	75	75	70	85	78	85	20	76.14	7.35

Note: SD: Standard Deviation

Before giving the results of the formal inter-rater reliability computations, it might be useful to examine the range and standard deviation figures informally. For example, for Portfolio Sample 1, rater range was 40 points. In grading terms that might mean one rater gave this portfolio a grade of D and another a grade of A+. Between any seven raters on a given portfolio, the average range is approximately 25 points suggesting a wide degree of assigned merit to the same portfolio by different raters. The standard deviations show a similarly wide disparity between raters in the scores given to any one portfolio.

In order to find out the inter-rater reliability of the subjective criteria, a measure of inter-rater reliability was computed using the procedure outlined by Hatch and Lazaraton (1991, pp.533-535). In this procedure, all the ratings are correlated producing a Pearson correlation matrix. In the case of 7 raters this produces a matrix of 21 pairs of correlations. For statistical balancing, the Pearson correlations are converted into Fisher Z transformations and an average of the 21 transformed correlations is taken. Pearson correlations for the first grading session are given in Table 3 below.

Table 3

Pearson Correlations for the first grading session

		Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7
Rater 1	Pearson Correlation	1.000	.299	.351	.158	.414	.275	.047
	Sig. (2-tailed)	.	.565	.495	.765	.414	.597	.929
	N	6	6	6	6	6	6	6
Rater 2	Pearson Correlation	.299	1.000	.891(*)	.869(*)	.742	.488	.226
	Sig. (2-tailed)	.565	.	.017	.025	.091	.326	.667
	N	6	6	6	6	6	6	6
Rater 3	Pearson Correlation	.351	.891(*)	1.000	.721	.763	.267	-.133
	Sig. (2-tailed)	.495	.017	.	.106	.078	.609	.802
	N	6	6	6	6	6	6	6
Rater 4	Pearson Correlation	.158	.869(*)	.721	1.000	.393	.174	.120
	Sig. (2-tailed)	.765	.025	.106	.	.441	.741	.822
	N	6	6	6	6	6	6	6
Rater 5	Pearson Correlation	.414	.742	.763	.393	1.000	.779	.391
	Sig. (2-tailed)	.414	.091	.078	.441	.	.068	.443
	N	6	6	6	6	6	6	6
Rater 6	Pearson Correlation	.275	.488	.267	.174	.779	1.000	.860(*)
	Sig. (2-tailed)	.597	.326	.609	.741	.068	.	.028
	N	6	6	6	6	6	6	6
Rater 7	Pearson Correlation	.047	.226	-.133	.120	.391	.860(*)	1.000
	Sig. (2-tailed)	.929	.667	.802	.822	.443	.028	.
	N	6	6	6	6	6	6	6

* Correlation is significant at the 0.05 level (2-tailed).

In Table 3 inter-rater pairings are repeated twice. There are 21 natural pairings, and those that had significant inter-rater reliability are shown. As can be seen from the table only raters 2 and 3, raters 2 and 4, and raters 6 and 7 had highly correlated ratings. One pair of raters, raters 3 and 7 were negatively correlated.

The average of the 21 transformed correlations is part of the formula for inter-rater reliability:

$$r_{tt} = \frac{nr_{AB}}{1 + (n-1)r_{AB}}$$

In this formula, r_{tt} stands for the reliability of all the judges' ratings, *n* stands for the number of raters and r_{AB} is the average correlation of ratings of all raters. r_{tt} is transformed back into a Pearson correlation value and that value is checked in a table of Pearson Product Moment Correlations to determine the combined correlation of raters and the significance of this correlation (Hatch & Lazaraton, 1991, p. 533).

These computations yielded a Fisher value of .894, which transforms into a Pearson correlation value of .71. This means that the inter-rater reliability is marginal (Hatch & Lazaraton, 1991).

After the first scoring, instructors met in a focus group discussion to define and clarify criteria for rating the writing portfolios. Following this discussion, instructors agreed on five analytic criteria suggested by the researcher which they agreed to use in scoring the next set of writing portfolios in the second scoring session. These criteria were Content, Organization, Language Use, Vocabulary and Mechanics.

After having defined the five analytic criteria, instructors were asked to rank order their perception of the relative importance in scoring the writing portfolios of each of these analytic criteria. The instructors' perceived relative importance of each of the five analytic criteria is shown in the rank orders in Table 4 below.

Table 4

Rank Order of Portfolio Analytic Criteria Weights

	Content	Organization	Language Use	Vocabulary	Mechanics	Total
Rater 1	1	4	3	2	5	15
Rater 2	1	2	4	3	5	15
Rater 3	1	2	5	3	4	15
Rater 4	1	2	4	3	5	15
Rater 5	1	4	2	3	5	15
Rater 6	2	1	3	4	5	15
Rater 7	2	1	3	4	5	15

Kendal W was computed to see the correlation among all raters using these five factors. The indicated correlation was .71. To look at the significance of the Kendal W value, the result was converted to a Chi Square value and examined in the appropriate table. Although there were some variations in raters' ranking of the five analytic criteria, the overall correlation was significant (at p<.005).

This suggests considerable agreement of opinion in respect to the valuing of the analytic criteria in portfolio scoring. However, the values the raters assigned to the five analytic criteria are not shown in the ranking. In the actual grading the raters may consciously or unconsciously be influenced by their own personal criteria. Given the agreement on valuing of analytic criteria indicated above, it is perhaps, somewhat surprising that there were major differences in how raters scored the portfolios in the second scoring using these analytic criteria as scoring guides which will be revealed in the next section.

Research Question 2: Inter-rater Reliability of the Analytic Criteria

In the second grading session, the instructors were introduced to the new analytic criteria adapted from Jacob et al. (as cited in Hughes, 2003, p. 104). Instructors carefully analyzed the analytic criteria and agreed on the original weights. The weights of Jacob et al. analytic criteria, as agreed upon by raters, analytic criteria are given in Table 5 below.

Table 5

The new analytic criteria weights

Criteria	Weights
Content	30
Language Use	25
Organization	20
Vocabulary	20
Mechanics	5

The scores of the instructors given to the other six new writing portfolios, using the analytic criteria, were collected. The scores are shown in Table 6 below.

Table 6

Portfolio grading with analytic criteria

Portfolio Sample	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Range	Mean	SD
7	65	71	69	70	73	85	76	20	72.71	6.39
8	85	80	77	82	83	96	98	21	85.85	8.02
9	80	79	77	76	83	85	100	24	82.85	8.19
10	45	70	70	72	70	93	86	48	72.28	15.15
11	80	64	67	63	79	79	95	32	75.28	11.44
12	80	80	80	69	65	85	89	24	78.28	8.47

Note: SD: Standard Deviation

These scores indicate that there is notable discrepancy among the instructors' grades. To confirm this, it again is useful to examine the range and standard deviation figures. For example, for Portfolio Sample 10, rater range was 48 points. In grading terms, that might mean one rater gave this portfolio a grade of F and another, a grade of A-. Between any seven raters on a given portfolio, the average range is again approximately 25 points suggesting a wide degree of assigned merit to the same portfolio by different raters. The standard deviations show a similarly wide disparity between raters in the scores given to any one portfolio.

In order to find out the inter-rater reliability of the analytic criteria a measure of inter-rater reliability was computed using the same procedure outlined by Hatch and Lazaraton (1991, pp.533-535). Pearson correlations for the first grading session are given in Table 7 below.

Table 7

Pearson correlations for the second grading session

		Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7
Rater 1	Pearson Correlation	1.000	.428	.499	.164	.481	-.272	.623
	Sig. (2-tailed)	.	.397	.314	.757	.334	.602	.186
	N	6	6	6	6	6	6	6
Rater 2	Pearson Correlation	.428	1.000	.960(**)	.732	.057	.454	.336
	Sig. (2-tailed)	.397	.	.002	.098	.915	.365	.515
	N	6	6	6	6	6	6	6
Rater 3	Pearson Correlation	.499	.960(**)	1.000	.563	-.041	.313	.428
	Sig. (2-tailed)	.314	.002	.	.245	.939	.546	.397
	N	6	6	6	6	6	6	6
Rater 4	Pearson Correlation	.164	.732	.563	1.000	.461	.815(*)	.366
	Sig. (2-tailed)	.757	.098	.245	.	.358	.048	.475
	N	6	6	6	6	6	6	6
Rater 5	Pearson Correlation	.481	.057	-.041	.461	1.000	.076	.668
	Sig. (2-tailed)	.334	.915	.939	.358	.	.885	.147
	N	6	6	6	6	6	6	6
Rater 6	Pearson Correlation	-.272	.454	.313	.815(*)	.076	1.000	.062
	Sig. (2-tailed)	.602	.365	.546	.048	.885	.	.906
	N	6	6	6	6	6	6	6
Rater 7	Pearson Correlation	.623	.336	.428	.366	.668	.062	1.000
	Sig. (2-tailed)	.186	.515	.397	.475	.147	.906	.
	N	6	6	6	6	6	6	6

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

In Table 7 inter-rater pairings are shown twice. There are 21 natural pairings and those that had significant inter-rater reliability are shown. As the table indicates, only raters 2 and 3 and raters 4 and 6 had highly correlated ratings. One pair of raters, raters 1 and 6 were negatively correlated.

The average of the 21 transformed correlations is part of the formula for inter-rater reliability:

$$r_{tt} = \frac{nr_{AB}}{1+(n-1)r_{AB}}$$

These computations yielded a Fisher value of .880, which transforms into a Pearson correlation value of .70. This means that the inter-rater reliability is, again, marginal (Hatch & Lazaraton, 1991).

It is interesting to see that the discussions of the new analytic criteria and the agreement on its sub-components did not make much difference in the inter-rater reliability when compared to the first grading session. Although the instructors seemed to agree on the original weights of the Jacobs et al. (1981) analytic criteria, it can be concluded that instructors still had their own subjective criteria in mind while scoring the portfolios in the second grading session.

Results of the Focus Group Discussions

Focus group discussions were held before and after the two grading sessions in order to have the seven participants discuss the portfolio assessment in their institution and express their opinions on how the criteria were used. Three focus group discussions were held. The number of questions in the focus group discussions differed according to the content of the discussions. The first focus group discussion consisted of nine questions, the second had five questions and the third had three questions (see Appendix D). This section presents the results of the data collected and analyzed to provide answers to the research questions.

Analysis of the Focus Group Discussions

The data gathered from the focus group discussions with the teachers were analyzed qualitatively through categorization and coding. The categories were mainly based on the research questions as well as teachers' perceptions of the portfolio implementation discussed after each scoring session. The analysis of the data revealed that teachers had similarities in their perceptions of portfolio implementation, but differences in their sense of importance of the five analytic criteria. The results of the focus group discussions will be presented under three headings: first focus group discussion, second focus group discussion and third focus group discussion.

Research Question 3: Instructors' General Perceptions of the Portfolio Assessment Scheme in Their Institution

First Focus Group Discussion

The first focus group discussion was held on the first day of the grading sessions. This discussion was a warm-up session on portfolios, their contents and importance and on the more general question of portfolio assessment. The researcher asked instructors questions on issues such as the implementation of portfolio assessment, the 5% value of portfolios in the overall grades of students, the effects of portfolio assessment on students' writing abilities and performance, and questions on portfolio contents and their suggestions as to various writing products that might be included in the future.

The results of the first focus group discussion indicate that the instructors are pleased with the implementation of portfolio assessment in writing classes. The positive opinions that the instructors mentioned could be grouped under two headings: 1) portfolio assessment has a positive impact on their instruction, 2) portfolios have positive effects on students.

The comments from the first focus group discussion related to these two headings are presented below.

1) Portfolio assessment has a positive impact on their instruction

Instructor 1: Teaching has become more organized and effective.

Instructor 2: Portfolio assessment has been administered for three years. It is a system of control by the teacher and revision by the student. It is an output of the education in which you also check what has been taught.

Instructor 3: You are able to see what you have taught.

Instructor 5: It allows drafting, double-checking, finalizing and in-class writing.

Instructor 7: Teachers get to know the students better…their potential and abilities. Teachers are able to observe the progress better.

The views mentioned in the first focus group discussion emphasize the importance of instructor and student working together. They mentioned that instructors and students are able to see learning progress better. The instructors also highlighted the idea that portfolio assessment is a system of control and feedback for the student, the teacher and the administration.

2) Portfolios have positive effects on students

Instructor 1: It promotes feedback to students and it results in motivation. In previous years, the writing classes were more teacher-centered…now students have more responsibilities.

Instructor 6: Portfolios also prepare the students for writing exams…so they feel more self-confident.

Instructor 7: I can clearly observe that students have more self-confidence during the exams

The above sentences taken from the first focus group discussion indicate that teachers mostly agree on the advantages of the portfolio assessment. Portfolios are considered to promote motivation through producing focused writing and taking on responsibility. Increased self-esteem is also another positive result of portfolio implementation as it allows students to more fully realize their abilities and to perform better on writing items in timed-exams.

During the first focus group discussion several suggestions arose on the following two issues: 1) The content of the portfolios could be elaborated 2) The weight of the portfolios could be increased.

The suggestions from the first focus group discussion related to the portfolio content and portfolio weight issues are given below.

1) The content of the portfolios could be elaborated

Instructor 1: Free writing can be introduced to students… there is a lack of this type. We are quite book-centered.

Instructor 2: Reflection papers are important; they help the student to view his own progress.

Instructor 4: More types of essays should be included… argumentative, opinion, compare and contrast essays. We should consider the level and the needs of the students in choosing contents… this university is not an English-medium university…

Instructor 7: Reports can be added… they need to collect data, analyze and reach to a conclusion.

Instructor 5: Reports seem a little bit advanced for our students. Before that we had better not teach writing 'letters' because students only memorize the format…

The instructors suggested that the content of the portfolio should be elaborated with more writing genres. Although Instructor 1 emphasized that free writing practices should be included in the portfolios, Instructor 4 insisted that free writing should depend on the level of the students, in that the beginning level students may have difficulties in that genre. Reflection papers on the other hand were highly recommended by Instructor 2, since reflection papers allow learners to "learn about learning" and get engaged in self-reflection (Paulson, Paulson & Meyer, 1991). Instructor 5 and 7 had opposing views on including reports. While Instructor 7 believed that report writing could promote some meta-cognitive skills, instructor 5 viewed reports as unrealistic items in terms of student levels. Including argumentative, opinion, compare and contrast essays and reports were also mentioned in the discussion as additional possible portfolio items in the future.

2) The weight of the portfolios can be increased

Instructor 1: The 5% can be increased to 10% or 15%…because writing is the most productive skill…one of the most effective courses.

Instructor 2: Yeah, I agree (with Instructor 1), it is something technical and can easily be increased.

Instructor 3: They seem to enjoy this portfolio process, so it can be increased.

Instructor 6: I think it should increase

Most of the instructors stated that the weight of the writing portfolios in the overall student grade should be higher. They asserted that 5% weighting of writing portfolios could be increased to 10% or 15%. They regarded writing as the most productive course in the program. Therefore, increasing the weight of portfolios was considered to provide additional positive impact related to the issues previously discussed.

Research Question 4: Instructors' Perceptions of the "Traditional" Writing Portfolio Assessment Criteria (Scoring 1)

Second Focus Group Discussion

In the second focus group discussion, which was held after the first grading session, instructors were asked questions about what goes into the grading in portfolio assessment, their perspectives on their own portfolio assessment criteria and problems they had that affected the reliability of the assessment.

The results of the second focus group discussion presented different aspects in the assessment of portfolios that are initially important for the instructors. The second focus group discussion can be grouped under two headings: 1) the criteria in assessing portfolios 2) the problems the instructors had that affected the reliability of the assessment.

The comments from the second focus group discussion related to these two topics are presented below.

1) The criteria in assessing portfolios

Instructor 1: It depends on the genre; in a letter format and organization is more important. It is important to see whether the student revealed the message or not.

Instructor 2: The neat and careful design of the portfolio is very important. Looking at the portfolio in general, whether the student has kept the quality high from the beginning to the end is important for me.

Instructor 3: Conveying the message through paragraph organization, planning and coherence… mechanical expectations…

Instructor 5: I look for what I taught to them; the evidence of progress between the first and second drafts.

Instructor 6: The portfolio itself is more important than the drafts included.

Instructor 7: The evidence of students' learning from their errors and how they applied what they have learned.

The responses given by the teachers clearly exemplify the differences among the instructors on the important aspects of assessing writing portfolios. Instructors 1 and 3 focused on the organization and the format of the genres and how the student conveys the message through that organization. Instructors 5 and 7 agreed on the evidence of students' learning from errors and using what has been taught as being the major criteria in their portfolio assessment measures. Mechanical considerations and the neatness of the portfolios are important considerations of instructors in the assessment of portfolios. Instructor 2 mentioned the appearance of the portfolio through-out the discussions as a main concern. Instructor 6 seems to agree with Instructor 2 in taking the portfolio into consideration as a whole.

These differences among the responses of the teachers may be an indication of why there was again only minor agreement in the second grading session. As mentioned before, in the ranking of composition sub-components during this discussion, the instructors' perspectives on these criteria differed significantly. Despite the high correlation among raters as to the relative importance of the five analytic factors (shown in Table 4) and despite the fact that the raters "agreed on" the relative weights given in the analytic criteria (shown in Table 5), the focus group discussions indicated there was much less real agreement on the criteria and their weighting than the quantitative data indicated.

2) The problems the instructors had that affect the reliability of the assessment

Instructor 1: I do not accept portfolios which exceed the deadlines.

Instructor 2: A good portfolio organization, presentation, good handwriting… we are trying to overcome this… you read very bad samples… when you read a good sample you are impressed… but it doesn't effect the overall grade much.

Instructor 3: I don't take these into consideration. The student gets the portfolio grade but my evaluation of the neatness and so on affects my own percentage on student's overall grade.

Instructor 5: I reduce grades from portfolios which exceed the deadlines…

Instructor 7 Handwriting affects the teacher. One first does not want to read a bad handwriting… but then you focus on meaning.

The sentences above taken from the second focus group discussion indicate several aspects that affect the reliability of instructors' scorings. While Instructors 2 and 7 stated that they were affected by the neat handwriting or the well-organized portfolio. Instructor 3 added that they did not affect the portfolio grade. Instructors 1 and 5 mentioned that they gave more importance to meeting the deadlines. Instructor 1 did not even accept late portfolios. (This yields a grade of F [Fail] in the student's 5% portfolio value). These differences again indicate the various aspects that instructors take into consideration during the assessment of writing portfolios.

Research Question 5: Instructors' Perceptions of the Analytic Criteria (Scoring 2)

Third Focus Group Discussion

The final focus group discussion was held at the end of the second grading session using the analytic criteria. During this grading session instructors discussed and agreed on the weights of the analytic criteria per Jacob et al. (as cited in Hughes, 2003, p. 104). The agreed-upon analytic criteria weights were: Content 30%, Language Use 25%, Organization 20%, Vocabulary 20% and Mechanics 5%.

The results of the third focus group discussion are grouped under three topics: 1) the instructors' perceptions of the analytic criteria, 2) the instructors' views on the comparison of the two scales used, 3) the instructors' suggestions about portfolio criteria in general. The first and the second topics provide response to the fifth research question.

The expressions from the third focus group discussion related to these three headings are presented below.

1) The instructors' perceptions of the analytic criteria

Instructor 1: Any criteria depend on the nature of the genre that I teach. Mechanics would weight more in a letter format…or in an essay content can weight 40.

Instructor 2: The criteria seem ok…sufficient…the weights may change according to the writing type.

Instructor 3: We had concrete data; it was good to have score ranges. Actually it was easier for me to score.

Instructor 4: It took so much time. I couldn't decide where to look at in the criteria. I don't think it is appropriate for portfolio assessment.

Instructor 5: The criteria caused me to give lower grades than I usually do, so I changed the scores I gave.

Instructor 7: Since we had criteria, it worked more reliable and objective approach.

The responses of the instructors on their perspectives of the analytic criteria also vary, although they had agreed on the weights of the new analytic criteria at the beginning of the session. Instructors 2, 3, and 7 were satisfied with the way the criteria were designed. They stated that the criteria reflected a reliable and objective approach. However, Instructors 4 and 5 were not comfortable with the criteria in terms of portfolio assessment. They believed that portfolio assessment should not be so limited by strict criteria. Instructor 1 also mentioned that each genre needs different criteria, so the criteria need to be flexible.

2) The instructors' views on the comparison of the two scales used

Instructor 1: The papers were read and scored in a more detailed way.

Instructor 2: In the second session, we had a compass…a guide…we scored according to that. In the first session, we made an evaluation depending on our school.

Instructor 3: Today's session was more reliable because the criteria was detailed

As the results indicate, three of the instructors above agreed that there was a significant difference in the two ways portfolios were assessed. They mostly agreed that a scoring guide and detailed criteria increased the reliability of the assessment.

3) The instructors' suggestions about portfolio criteria in general

Instructor 2: These criteria should be often revised and rearranged… Otherwise it keeps the teacher gradually away from the assessment procedures.

Instructor 1: Yeah, I agree that criteria should be flexible.

Instructor 4: A scale of 1-6 would be better to avoid this many ranges in the grades. When portfolios are considered I do not look at these scores and such detailed criteria.

Instructor 5: The rating of the two raters can be accepted only if the product, the portfolio itself is scored. If the process is important in scoring, then one rater—the writing teacher—should be assessing the portfolios. Because the teacher can see the progress…knows the student better.

Although there was not a strong degree of inter-rater reliability in the scores in both grading sessions, most teachers agreed on the desirability of having criteria. However, they insist that the criteria should be flexible and open to change according to written genre and over time. Having two raters was an interesting issue raised by Instructor 5 in this discussion. She stated that if "process" evaluation is the main concern of the portfolio assessment, then only the writing teacher should assess the portfolios. However, if the whole "product" is of primary concern, then two raters can judge. This uncertainty about the main purpose of portfolio assessment criteria tends to indicate that insufficient information is given to instructors by the institution about the portfolio goals and objectives.

Results of the Interviews

In this section results of the interviews will be discussed. Participants were interviewed in order to get information about their perceptions of portfolio assessment in the institution and on the criteria used in the two portfolio grading sessions. The interviews were held individually after the grading sessions were over. The interviews consisted of six questions, with the last question having three sub-questions (see Appendix C). This section presents the results of the data collected and analyzed to provide answers to the original research questions.

Analysis of the Interviews

The data gathered from the individual interviews with the teachers were analyzed qualitatively through categorization and coding. The categories were mainly based on the research questions as well as teachers' perceptions of the portfolio implementation at their institution and their suggestions for future portfolio assessment. The analysis of the data revealed that teachers had similarities in their perceptions of portfolio implementation, but differences in their criteria preferences and suggestions for the future. The interview results will be presented under four headings: instructors' general perceptions of portfolio implementation, instructors' perceptions of the assessment criteria currently used in the institution, instructors' perceptions of the analytic assessment criteria and instructors' suggestions for future applications.

Research Question 3: Instructors' General Perceptions of Portfolio Implementation in the Institution

Interview results about the portfolio implementation will be considered under these three categories: 1) The positive sides of portfolio implementation, 2) other aspects of portfolio implementation, 3) training of teachers in the portfolio process.

The expressions from the interviews related to these three headings are presented below.

1) The positive sides of portfolio implementation

The instructors expressed the benefits of the portfolio assessment in their institution in the following ways.

Five participants out of seven stated the positive effect of portfolios on the evaluation of students. The below sentences are examples taken from the interviews.

Instructor 1: Comparing it to the previous portfolio-free program, I can know the student and grade his performance better.

Instructor 2: This is the system that we gain information about students the best. We have the data that we can evaluate.

Instructor 3: Students are more prepared for the writing exams in this way and we are able to observe their abilities better.

Instructor 6: We can observe how far the students developed between the first and the last assignments.

Instructor 7: You can get to know your students better. You know their capacities; therefore, you can catch the clues of plagiarism or the help of a proficient English user in student work.

The sentences above indicate that portfolios provide the instructors with accurate information about students' writing abilities. Moreover, three instructors stated that the evaluation of students' writing performance has become fairer and better because students are able to indicate what they can do both in the portfolios and the writing exams.

Two instructors out of seven stated that the portfolio is the product of both the teacher and the student. The below sentences are examples taken from the interviews.

Instructor 2: I think it is two-sided; on one hand the student creates something in the text format in a foreign language, on the other hand, the teacher can observe what has been taught.

Instructor 4: It is both my and the student's product because I seek what I taught in the classroom.

As the sentences above indicate, two instructors emphasize that portfolios supply a source of evaluation of the teacher and the student. It not only assesses the writing abilities of the students but also provides an opportunity for teachers to evaluate themselves.

Six instructors mentioned the positive relationship between portfolio assignments and the writing exams. The sentences below are taken from transcriptions about this issue.

Instructor 1: Students are more prepared for the class.

Instructor 2: Portfolios give the students a sense of homework.

Instructor 3: It depends on the students; for some it is a study of preparing assignments and getting scores out of that.

Instructor 4: It prepares the students for the exam…shows writing abilities depending on the assignments.

Instructor 5: It is an evidence of what has been done through the year…prepares the student to the exam.

Instructor 6: It is a compilation of student assignments and therefore represents student growth.

All instructors above emphasized that portfolios are tools to have students prepare for the exam and classroom activities in a more organized and conscious way. These statements illuminate the positive force of portfolios on students' assignment responsibilities.

Only one instructor mentioned the motivational effect of the portfolios. Instructor 1 stated that:

> The portfolio assessment should have been practiced before. Students are very much involved in the learning process. I believe it is positive for student motivation.

Although the other positive statements about portfolios imply the motivational aspects of portfolios, it is interesting that only one instructor specifically commented on the motivational effect of portfolio assessment during the interviews.

2) Other aspects of portfolio implementation

The instructors commented on some other aspects of the portfolio assessment in their institution. These aspects consist of issues such as giving feedback to students and students' freedom of choice on portfolio contents. The comments were given in the following ways.

All interviewees stated that students gain understanding of their writing primarily by direct feedback from their teachers. The below sentences are examples taken from the interviews.

Instructor 1: I give them feedback.

Instructor 2: They mostly get feedback from me.

Instructor 3: They've always got feedback from me.

Instructor 4: From me. I always use portfolio conferences.

Instructor 5: From me.

Instructor 6: I give the feedback.

Instructor 7: The teacher gives the feedback.

As seen from the sentences above students primarily get feedback about their writing from their teachers. Instructor 4 was the only person who mentioned the conference method to give feedback on writing.

Three instructors out of seven stated that they also use peer feedback in their writing classes. The below sentences are examples taken from the interviews.

Instructor 2: Portfolio practices can turn into group works and students can learn through the communication between themselves and their friends. I don't see portfolios as a system of thoroughly teacher feedback. Students gain knowledge through in-class discussions and see what they lack.

Instructor 5: …very occasionally there is peer feedback.

Instructor 6: They have access to each other's portfolios and they give informal feedback to each other.

The above raters stated that although they are the primary source of feedback, they try to use peer feedback in their classrooms. However, these peer feedback practices do not seem to be a systematic and formal assessment type. They are occasional and limited.

Among other aspects of portfolio implementation, teachers were also asked about the freedom of topic choice students should have. The instructors expressed their ideas about student freedom of topic choice in the following ways.

Six interviewees out of seven stated that the syllabus factor enables the students to have little choice in portfolio contents, which they should have. The below sentences are examples taken from the interviews.

Instructor 1: There is no freedom of choice, it is a totally teacher centered and syllabus-based system. I'd rather students choose at least one project themselves.

Instructor 2: We have to follow a book and our portfolio system is set on books. So student views are not considered much. There could be some items included chosen by the student.

Instructor 3: Until now they've written on topics already decided by the teachers. There should be some freedom so I try to elaborate the topics according to their interests.

Instructor 5: I believe there should be freedom. I elaborate the topics or writing types with my students, depending on the classroom atmosphere.

Instructor 6: This year I let my students write on the topics they like. In this way they are more motivated and enthusiastic to work.

Instructor 7: I give them various alternatives on relevant topics. They have freedom to choose out of those.

As noted above by the instructors there are also differences in the implementation of the writing courses in the classrooms. Some students are free to choose their topics, whereas others are more teacher-controlled. Only one instructor stated that there should not be freedom of choice for students in the following way.

Instructor 4: There is no freedom of choice and there shouldn't be. I believe we should be deciding the content.

It is interesting to see that a portfolio assessment method, which nominally promotes learner responsibility and autonomy, stands in some contradiction to practices which totally reject student freedom of choice.

3) Training of teachers in the portfolio process

Instructors were also asked whether they had had or needed any training on portfolio assessment. Except for one instructor all the interviewees revealed that they need adequate training on portfolio assessment in the following ways.

All participants stated that they did not have a formal training on portfolios. The below sentences are examples taken from the interviews.

> Instructor 1: I didn't have any training, only a general explanation. I do believe that education is needed…not maybe in the implementation, but in the assessment of portfolios.
>
> Instructor 2: We didn't have formal training in the institution, but we join seminars and try to get help from more experienced people.
>
> Instructor 3: I didn't have any training. I only had some information from the writing coordinator, there is a standard approach to portfolios and I try to practice that in my classrooms.
>
> Instructor 4: I didn't have any training, but I read a lot and asked my friends. I don't think we need training; it is something that naturally generates at school.
>
> Instructor 5: No, I didn't, but we help each other.
>
> Instructor 6: I didn't have any, but I think we should. There aren't many meetings on writing but I think there should be a meeting on each item in the portfolio.

Instructor 7: I got some help from the writing coordinator in the beginning. But not any formal training has taken place. I would be nice to learn about the new applications in the field.

As seen in the statements above, none of the teachers had adequate, preplanned or formal training for the portfolio assessment implementation in their institution. However, they all had several sources of information. The information sources they mentioned were seminars, the writing coordinator, more experienced peers, and articles. Except for Instructor 4, they all have positive attitudes towards having formal training.

Research Question 4: Instructors' Perceptions of the "Traditional" Writing Portfolio Assessment Criteria Currently Used in the Institution

Four instructors out of seven stated that they did not think that there is consistency among teacher evaluations of student portfolios. The below sentences are examples taken from the interviews.

Instructor 1: I don't think there is consistency because we are not given criteria. It is totally left to the teachers. As seen in the previous sessions, if each teacher has different criteria, we can't reach standardization.

Instructor 5: No, there can be significant differences among teachers' scores. You are evaluating effort. Student is very important. If you know the capacity of a student and see that the work is appropriate for his capacity, you grade it fairly. But if you think that the student produced below capacity, you grade it differently.

> Instructor 6: No, I don't think there is consistency. Some teachers take the classroom participation of a student into consideration and give high scores in portfolios although the portfolio is not worth a high grade. For some teachers grammar can be more important whereas vocabulary is important for others.
>
> Instructor 7: There isn't much consistency. Some teachers see this as an opportunity to help the students raise their overall grades.

As the statements above indicate, these instructors have observed inconsistencies among teacher evaluations of portfolios. What they commonly mention is the fact that each teacher has different criteria while assessing portfolios. These criteria differ in terms of ESL writing sub-components, classroom participation or emotional relationship between the teacher and the student.

Instructors 2, 3, and 4 stated that they believe there is consistency among teacher evaluations of student portfolios. The below sentences are examples taken from the interviews.

> Instructor 2: Teachers are encouraged to use their initiatives and through time I believe consistency develops.
>
> Instructor 3: I think there is, especially among teachers who have given writing courses for a long time. There can be consistencies among novice and experienced teachers too as long as the novice ask for help from the experienced.

Instructor 4: There is consistency between my colleagues and me with whom I share the same office. We are always informed about how we grade the portfolios and we always discuss our views and experiences.

The above sentences taken from the interviews show that teachers who believe there is consistency among raters' grades seem to be limited to more or less specific cases. Consistency seems to be either a process that will be developed in time or limited to a number of people who share physical office space with others.

Research Question 5: Instructors' Perceptions of the Analytic Assessment Criteria

When asked about their perceptions of the portfolio assessment criteria, instructors gave different responses. The responses mostly differed to the extent that they felt analytic criteria match their own subjective criteria.

Three instructors out of seven stated that the analytic criteria matched their own criteria and revealed positive feelings about it. The sentences below are samples from the interviews.

Instructor 3: It matches my criteria in all ways. The criteria limited me but helped to discriminate the 'good' and 'bad' paper.

Instructor 6: The method is nice and it matches my subjective criteria.

Instructor 7: Yes it does.

Two instructors out of seven stated that the analytic criteria partially matched their subjective criteria and mentioned some positive perceptions about the criteria.

Instructor 1: It mostly fits my criteria. I'd rather weight content and organization same.

> Instructor 2: It doesn’t totally match my criteria but I can say 40% of them match. However, criteria or limitations are not bad, they prevent disorder. I find them positive.

As seen from the sentences above the teachers who said the analytic criteria both totally and partially matched their subjective criteria were satisfied with the criteria. They believed that through the criteria, ‘good’ writing papers were considered more noticeable and got fairer grades. They also agreed on the notion of criteria or standardization to prevent chaos in the assessment.

Instructors 4 and 5 stated that the analytic criteria did not match their subjective criteria in any way.

> Instructor 4: It didn’t match at all. It had so many criteria that I couldn’t decide which one to look at. I don’t think that standard criteria can work in portfolio assessment. Criteria shouldn’t be standard or universal. I feel like somebody is interfering my business.
>
> Instructor 5: No, it didn’t match. I can’t deal with grades as in those criteria. I realized that I had given very low grades, and then modified my grades. I prefer giving an overall grade at first glance. After that, I divide it into sub-categories.

The responses of instructors 4 and 5 focused on the detailed and standardized characteristics of the analytic criteria. One of the reactions against the criteria deals with the sub-components of the analytic criteria and how those sub-components were divided into sub-grades. The other reaction they gave was to the criteria itself.

Instructors 4 and 5 seem fond of having no criteria and grading the portfolios impressionistically in the way they always do.

Instructors' Suggestions for Future Applications

This section gives the instructors' suggestions for future portfolio assessment practices. Their suggestions focus on whether there should be standard criteria or not in portfolio assessment. Thus, the headings in this section will be 1) Instructors who want change in the assessment of portfolios and 2) Instructors who do not want change in the assessment of portfolios.

1) Instructors who want change in the assessment of portfolios

Three instructors out of seven stated that they prefer a change in the assessment of portfolios. The below sentences are examples taken from the interviews.

> Instructor 1: I think the evaluation of portfolios should change. And this change should depend on the need and students' demands. Each time we will have different types of students and I believe grading should change according to this. Criteria of which the sub-components are flexible depending on the genre would be good.
>
> Instructor 2: Everything is changing. Five years from now the grading system will change too. But speaking of today, we need to assess these depending on some criteria. However, the criteria should be flexible according to subject matter.

> Instructor 3: In order to be more objective we should have criteria. And I am also on the side of having two raters grade the portfolios.

The suggestions above, given by three instructors, mainly favor a change in the portfolio assessment in terms of having criteria. Instructors 1 and 2 propose having criteria in portfolio assessment as long as these can be flexible according to the student profiles and capacities and writing genres. Instructor 3 raises the issue of objectivity and provides solutions by offering having criteria and having two raters score the portfolios.

2) Instructors who do not want change in the assessment of portfolios

Four instructors out of seven stated that they would like to grade portfolios the way they have always done. The below sentences are examples taken from the interviews.

> Instructor 4: I am happy with my own criteria and method of scoring. I think we should be emotional toward a product produced by a student. We shouldn't be too strict.
>
> Instructor 5: I do not have any problems with my own method. I also take into consideration the student's other grades that he got from exams.
>
> Instructor 6: I'd like to grade them the way I always do because even if there are criteria, I am sure each teacher will grade them according to the criteria they give more importance to.
>
> Instructor 7: I am happy with my subjective criteria, but I'd like to follow the new things in the literature.

As the statements above demonstrate more than half of the instructors stated that they were comfortable with the way they assess writing portfolios. Only instructor 6 had doubts about an acceptance of the new criteria. It is mentioned that instructors will be eager to use their initiatives and their subjective criteria even though the criteria are modified.

Conclusion

In this chapter, the data collected from portfolio grading sessions, focus group discussions and interviews were analyzed and interpreted. The results will be further exemplified in the next chapter.

CHAPTER V: CONCLUSION

Overview of the Study

This study investigated the inter-rater reliability of the portfolio assessment criteria currently in use and the new portfolio assessment criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department. The study also aimed to learn the perceptions of the instructors on the portfolio assessment implementation, the portfolio assessment criteria currently used and the new portfolio assessment criteria proposed.

This study addressed the following research questions:

1. What is the inter-rater reliability of Basic English teachers using the "traditional" writing portfolio assessment criteria prescribed at Yıldız Technical University, School of Foreign Languages, Basic English Department?
2. What is the inter-rater reliability of Basic English teachers using the new writing portfolio assessment criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department?
3. What are the instructors' general perceptions of the writing portfolio scheme at Yıldız Technical University, School of Foreign Languages, Basic English Department?

4. What are the instructors’ perceptions of the use of writing portfolio assessment criteria presently used at Yıldız Technical University, School of Foreign Languages, Basic English Department?
5. What are the instructors’ perceptions of the use of writing portfolio assessment criteria proposed for Yıldız Technical University, School of Foreign Languages, Basic English Department?

In order to fulfill the purposes of the study, three sets of data were collected: Instructors’ portfolio scores assigned to 12 student portfolios using both sets of the assessment criteria, results of three focus group discussions, and the results of the teacher interviews. The participants were seven writing instructors currently working at Yıldız Technical University, School of Foreign Languages, Basic English Department. The participants attended the first portfolio grading session and scored 6 portfolios using their subjective, traditional criteria. In the second grading session the participants scored another 6 portfolios using the new analytic criteria. The participants were asked to express their opinions about both of the criteria and the portfolio assessment implementation, in general, in the focus group discussions and the individual interviews. The first and the second focus group discussions were held during the first portfolio grading session. The final focus group discussion was held after the second portfolio grading session. Individual interviews took place after this on the same day.

The data were analyzed in three stages. First, the instructors’ portfolio scores given in two grading sessions using both of the assessment criteria were analyzed for inter-rater reliability using Pearson Correlations and Fisher Z Transformations. Second, focus group discussions were transcribed, categorized and coded according to the purpose of the study and the research questions. Finally, individual interviews were transcribed, categorized and coded focusing on the purpose of the study and the research questions.

In this chapter, the major findings of the study will be summarized and discussed. The chapter will also present pedagogical implications drawn from the findings, the limitations of this study, and suggestions for further studies.

Discussion of Findings

This section discusses the major findings and the conclusions that have been drawn through the data collection process. The findings of the study will be presented in three different sub-sections referring to each research question: the inter-rater reliability of the subjective criteria, the inter-rater reliability of the analytic criteria, the instructors' perspectives on the portfolio assessment implementation in their institution, the instructors' perspectives on their subjective criteria and the instructors' perspectives on the analytic criteria.

The Inter-Rater Reliability Using the Subjective Criteria

The analysis of the results revealed that the inter-rater reliability for the subjective criteria was 0.71. Therefore, we concluded that the scores are only marginally consistent (Hatch & Lazaraton, 1991).

The Inter-Rater Reliability Using the Analytic Criteria

The analysis of the results revealed that the inter-rater reliability for the analytic criteria was 0.70. Therefore, we concluded that the scores are again, only marginally consistent (Hatch & Lazaraton, 1991).

The Instructors' General Perceptions of the Writing Portfolio Scheme in Their Institution

Analysis of the results concerning the instructors' general perceptions of the portfolio assessment implementation in their institution reveals that most of the instructors find this practice satisfactory. This satisfaction is based on the instructors' positive attitudes towards various characteristics of portfolio assessment such as having a positive impact on instruction and on students, assessing directly what is taught, assessing student performance fairly and accurately, encouraging student self-esteem and motivation, enabling students to see their development in writing skills, and providing the instructors with accurate information about students' writing abilities. The results support Hamp-Lyon and Condon's (2000) claim that portfolios provide a broad and accurate view of students' writing abilities.

However, the instructors reported different opinions in terms of specifics of portfolio assessment, such as the freedom of choice the students have on deciding the portfolio contents, training of teachers in the portfolio process, and the feedback types used in classrooms. The results of the interviews show that students had limited freedom of choice and the extent of that freedom was different in each writing class, depending on the instructors' inclination. Only one instructor was strictly negative about giving students choice and suggested that teachers should be the sole decision-makers of the content. In terms of training all instructors stated that they did not receive formal training on portfolios, but were often in communication with their colleagues about each other's practices. Except for one instructor, they all agreed on the desirability of having formal training. Three instructors out of seven stated that they used peer feedback additionally in their classrooms. Only one instructor mentioned using portfolio conferences as a form of feedback.

The results of the interviews and the focus group discussions also indicated that the majority of the instructors agreed that the content of the portfolio should be elaborated with more writing genres and that the weight of the writing portfolios in the overall student grade should be higher.

The Instructors' Perceptions of the "Traditional" Writing Portfolio Assessment Criteria

The overall picture of the results of the interview and focus group discussions with seven instructors showed that instructors had different criteria while assessing writing portfolios. Two instructors out of seven stated that they primarily focused on the organization and how the message is revealed. Another two stated that they looked for development through learning from errors. Two instructors agreed on the presentation of the portfolio as a whole. The majority of the instructors also mentioned mechanical considerations in grading the portfolios.

In the five analytic criteria rank order comparison—Content, Organization, Language Use, Vocabulary, and Mechanics—it was observed that raters had considerable agreement with the relative importance of the analytic criteria.

In terms of the instructors' beliefs about the reliability of teachers' portfolio scores in the institution, the results show that four instructors out of seven stated that there are inconsistencies among teacher evaluations of portfolios. There are several sources of these inconsistencies. Although instructors agreed on analytic criteria rank order in principle, each instructor might still assign different personal weights considering the ESL writing sub-components. Teachers' scores are also affected by students' classroom participation or the relationship between the teacher and the student. These beliefs about the inconsistencies in scoring are also supported by the problems that instructors stated that they believed affected the reliability of the portfolios. The valuing of the organization of the portfolios, handwriting, and meeting deadlines are some other aspects affecting reliability of scores. On the other hand, three instructors out of seven believe there is consistency among raters' grades. The possible sources of rating consistency were all informal. They stated that agreements among the colleagues who share the same ideas or office space contributed to consistency.

The Instructors' Perceptions of the Analytic Criteria

The analysis made to identify whether there are any differences between the perspectives of the instructors on the analytic criteria revealed both strongly positive and slightly negative tendencies towards the analytic criteria. Five instructors out of seven were satisfied with the way the new analytic criteria were designed and stated that the analytic criteria matched their own criteria. They believed that the criteria provided objectivity and reliability. They also agreed on the notion of criteria or standardization to prevent chaos in the assessment.

However, two instructors were neither pleased with the analytic criteria nor with having any criteria in the assessment of portfolios. The detailed sub-categories of the analytic criteria and the ranges of those sub-categories were considered to be too standardized and strict which they believe is not appropriate for portfolio assessment.

Although more than half of the instructors were satisfied with the analytic criteria, only three instructors favored a change in the portfolio assessment criteria in the institution. Among the suggestions given by these three instructors were that the criteria should be flexible according to the genre and that two raters should grade portfolios. Four instructors stated that they were comfortable with the way they assess writing portfolios. However, they added that the criteria need to be flexible and often revised with a view that a change in the assessment criteria might be needed. Only one instructor among these four did not want a change in the assessment of portfolios because the instructor believed that any assessment criteria would be overshadowed by teachers' use of their own subjective criteria.

Pedagogical Implications

According to the results of the study, some sort of analytic criteria will be recommended for Yıldız Technical University, School of Foreign Languages, Basic

English Department. Although in both of the grading sessions the results of the inter-rater reliability were identical, as Williams (1998, as cited in Song & August, 2002) argues that standardization is needed especially in performance assessments. He adds that without standards for implementation and outcomes, portfolio assessment will be unfair because it increases the subjectivity teachers bring to evaluation. Some recommendations can be made to improve the analytic criteria. Since some raters stated that the criteria were too detailed that some other criteria, simpler or fewer ones should be considered. Moreover, an assessment framework that addresses the longitudinal dynamic aspects of the evaluation heuristic such as that developed by Hamp-Lyons and Condon (see p. 47) can be formulated. Their dimensions for assessing portfolios include observation of developmental processes in the characteristics of the writer, characteristics of the portfolio as a whole, characteristics of the individual texts and of the intratextual features. In contrast the Jacobs et al. (1981) analytical scale is primarily used for final product grading. Some of the instructors suggested, and the literature supports the view, that portfolio assessment should include assessment of process as well as assessment of the portfolio products as a whole. The Hamp-Lyons and Condon proposal incorporates these process dimensions which could be more appropriate in terms of the overall purpose of portfolio assessment. The application of these dimensions should be a part of our departmental discussions in the future.

The results of my study will be shared with our teachers. Some of them may be shocked at variability of ratings given to the same portfolio and try to arrive at a more consistent system of scoring.

Lacking some sort of formal criteria, it is recommended that more portfolio grading reliability might be attained if teachers involved in portfolio grading met in a discussion group before grading and reviewed several portfolios, mentioning how they might grade these and why.

The results of the study also indicated that instructors should be given professional training in order to be able to implement portfolio assessment more effectively and consistently. As Lumley and McNamara (1991) indicate, the training of raters is crucial in any testing condition. Moreover, the training sessions should include sample rating sessions and discussions afterwards. More writing instructors need to be involved in the discussion and design of the role of portfolios and criteria in portfolio assessment.

The active involvement of the students in their own language learning process can also be encouraged by giving students active roles in the decision-making process of portfolio assessment. As Paulson and Paulson (1994) recommend students can help set the standards, contents and the focus of the portfolios. This would result in better results in student self-reflection and self-monitoring.

It is also recommended that the 5% weight of portfolio assessment should be higher and other genres in portfolio construction need to be considered.

Limitations of the Study

One of the major limitations of the study is the number of participants which was limited to one third of the total number of writing instructors. Having more raters participate in the study could have helped reach more general results and brought further insights to the results of the study. This would also have informed instructors about the extent of our rating inconsistency and promoted some discussions as to how this could be improved.

Another limitation is about the analytic criteria. The inter-rater reliability might have been higher if the analytic criteria were developed on a more formal basis with more contribution from the raters. Finally, rater interactions may have been too limited due to time constraints. Since all the grading sessions, focus group discussions and interviews were held during workdays after school hours, the instructors had to stay for extra hours voluntarily. Due to this, the researcher had to keep the study within a reasonable time period which necessarily limited some of the discussion and planning which needs to take place.

Suggestions for Further Studies

In further studies, which look at the inter-rater reliability of raters, more data from more raters could be collected. Gathering information on students' perceptions would also be useful. Students' views of writing portfolios and assessment could be researched. Another research study might investigate the inter-rater reliability of raters within the analytic scale. The scores given to each category by a number of raters can be analyzed quantitatively. The sub-components on which raters have the lowest inter-rater reliability can be investigated and some suggestions for improvement can be made.

Finally, another research study, which looks at the inter-rater reliability of raters in the writing final exam, can be conducted. The scores given to each student by two raters can be analyzed quantitatively. The results of the study may provide useful information for the implementation of the writing final examination and assessment of the final papers.

Conclusion

This study investigated portfolios as an alternative assessment system to assess writing in the L2 classroom, as well as inter-rater reliability of teachers as writing evaluators at Yıldız Technical University, School of Foreign Languages, Basic English Department. The data was collected through writing portfolio scores from two grading sessions, focus group discussions and interviews.

The results revealed that the instructors have a positive attitude towards writing portfolio assessment, yet felt it can be improved by elaborating the content of portfolios, providing training sessions for the teachers and standardizing the assessment procedure of the portfolios.

The results of the quantitative data revealed that there is no real difference between the results of the two portfolio grading sessions (subjective and analytic) in terms of their level of inter-rater reliability. However, it is proposed that some sort of analytic criteria should be developed at Yıldız Technical University, School of Foreign Languages, Basic English Department in order to help establish standardization in the assessment. These criteria need to be further discussed in detail and, perhaps, simplified and/or modified with the contribution of instructors.

REFERENCES

Anderson, J. O., Bachor, D. & Baer, M. (2001). Using Portfolio Assessment to Study Classroom Assessment Practice. (ERIC Document Reproduction Service No. ED 462 445).

Association of American Publishers. *Standardized Assessment Primer.* www.publishers.org

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Barnhardt, S. Kevorkian, J. & Delett, J. (1998). *Portfolio Assessment in the Foreign Language Classroom*. National Capital Language Resource Center. (ED 448602)

Breland, H. M. (1996). *Writing Skill Assessment: Problems and Prospects. Policy Issue Perspective Series*. Princeton, NJ: Educational Testing Service.

Brindley, G. (2001). Assessment. In R. Carter & D. Nunan (Eds.). *The Cambridge Guide to Teaching English to speakers of other languages* (pp. 137-143). Cambridge: Cambridge University Press.

Brown, H.D. (2004). *Language Assessment: Principles and Classroom Practices*. NY:Longman

Brown, R. (1986). Evaluation and learning. In Anthony R. Petrosky and David Bartholomae (Eds.) *The Teaching of Writing* (pp. 114-130).

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ, USA: Prentice Hall Regents.

Brown, J. D. & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32, 4, 653-675.

Brown, J. D. & Rodgers, T. D. (2002). *Doing Second Language Research*. Oxford: Oxford University Press.

Bushman, L. & Schnitker, B. (1995). Teacher Attitudes on Portfolio Assessment, Implementation, and Practicability. (ERIC Document Reproduction Service No. ED 388 661)

Butler, P. (1997) Toward a definition of alternative assessment. In Priscilla Butler (Ed.) *Issues in Alternative Assessment: The Japanese Perspective.* Kwansei Gaikun University, Nishinomiya (Japan) Language Center, 1-10.

Cohen, A. D. (1994). *Assessing Language Ability in the Classroom*. USA: Heinle & Heinle Publishers.

Cole, D. J. Ryan, C. W. Kick, F. & Matthies, B. K. (2000). *Portfolios Across the Curriculum and Beyond*. California: Corwin Press.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching,Assessment*. Cambridge: Cambridge University Press.

Cumming, A. (2001). Learning to write in a second language: Two decades of research. In Rosa M. Manchon (Ed.) *International Journal of English Studies*. 1, 2, 1-23. (ED 465286)

Daiker, D. A. Sommers, J. & Stygall, G. (1996). The pedagogical implications of a college-placement portfolio. In Edward M. White, William D. Lutz and Sandra Kamusikiri (Eds.), *Assessment of Writing: Politics, Policies, Practices*, 257-270. New York: The Modern Language Association of America.

Demirel, Ö. (2004). *The ELP Project in Turkey*. Paper presentation in the 3rd Joint International ELT Conference at Trakya University, Faculty of Education, Edirne, Turkey.

Dinçman, P.S. (2002). Teachers' *Understanding of Projects and Portfolios at Hacetttepe University School of Foreign Languages Basic English Division.* Unpublished Master's thesis, Bilkent University: Ankara.

Douglas, D. (2000). *Assessing Language for Specific Purposes*. Cambridge: Cambridge University Press.

Ewing, S. C. (1998). Alternative assessment: Popularity, pitfalls, and potentials. *Assessment Update*, 10, 1, 1-12.

Gearhart, M. & Herman, J. L. (1998). Portfolio assessment: Whose work is it? Issues in the use of classroom assignments for accountability. *Educational Assessment*, 5, 1, 41-55.

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

Georglou, I. S., & Paulov, P. 2002. 'Language Portfolio'. *La Nuova Italia Editrice*. p.12-21. (http://www.oup.com/pdf/elt/it/nov02c.pdf)

Gottlieb, M. (2000). Portfolio practices in elementary and secondary schools: Toward learner-directed assessment. In G. Ekbatani & H. Pierson (Eds), *Learner directed assessment in ESL,* 89-104. Mahwah, NJ: Lawrence Erbaum.

Grabe, W. & Kaplan, R. B. (1996). *Theory & Practice of Writing*: *An Applied Linguistic Perspective*. UK: Longman.

Gronlund, N. E. (1998). *Assessment of Student Achievement*. MA: Allyn & Bacon A Viacom Company.

Gussie, W. F. & Wright, R. (1999). *Assessment of the Implementation of Portfolio Assessment in K-8 School Districts in New Jersey*. Paper presented at the Eastern Educational Research Association Conference, South Carolina. (ERIC Document Reproduction Service No. ED 429 996)

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In Liz Hamp-Lyons (Ed.) *Second Language Writing*, 241-266.

Hamp-Lyons, L. (1996). Second language writing: assessment issues. In Barbara Kroll (Ed.) *Second Language Writing,* 69-87.

Hamp-Lyons, L. & Condon, W. (2000). *Assessing the Portfolio*. NJ: Hampton Press, INC.

Hamp-Lyons, L. & Condon, W. (1993). Questioning assumptions about portfolio-based assessment. *College Composition and Communication*, 44, 2, 176-190.

Harper, A. E. & Misra, V. S. (1976). *Research on Examinations in India*. New Delhi: National Council of Educational Research and Training.

Hendrickson, J. M. (1984) The treatment of error in written work. In Sandra McKay (Ed.) *Composing in a Second Language*, 145-159.

Henning, G. (1993). Issues in evaluating and maintaining an ESL writing assessment program. In Liz Hamp-Lyons (Ed.) Assessing *Second Language Writing in Academic Contexts*, New Jersey: Ablex Publishing, 279-291.

Herman, J. L., Aschbacher, P. R. & Winters, L. (1992). *A practical Guide to Alternative Assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Herman, J.L., Gearhart, M. and Baker, E. L. (1993). Assessing writing portfolios: Issues in validity and meaning of scores. *Educational Assessment*, 1, 3, 201-224.

Hirvela, A., & Pierson, H. (2000). Portfolios: Vehicles for authentic self-assessment. In G. Ekbatani & H. Pierson (Eds), *Learner directed assessment in ESL,* Mahwah, NJ: Lawrence Erbaum, 105-126.

Hughes, A. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Hyland, K. (2003). *Second Language Writing*. Cambridge: Cambridge University Press.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Harthfield, V. F. & Hughery, J. B. (1981). *Testing ESL Composition: A Practical Approach*. Rowley, MA: Newbury House.

Karslı, E. S. (2002). *The Inter-rater reliability of Two Alternative Analytic Grading Scales for the Evaluation of Oral Interviews at Anadolu University School of Foreign Languages.* Unpublished Master's thesis, Bilkent University: Ankara.

Larson, R. L. (1996). Portfolios in the assessment of writing. In Edward M. White, William D. Lutz and Sandra Kamusikiri (Eds.), *Assessment of Writing: Politics, Policies, Practices*. New York: The Modern Language Association of America, 271-283.

Leeming, S. (1997) Incorporating alternative assessment in the teaching of EFL writing: Responding to multiple drafts. In Priscilla Butler (Ed.) *Issues in Alternative Assessment: The Japanese Perspective.* Kwansei Gaikun University, Nishinomiya (Japan) Language Center, 50-55.

Lumley, T. & McNamara, T. (1993). *Rater Characteristics and Rater Bias*. Paper presented at the Language Testing Colloquium. (ERIC Document Reproduction Service No. ED 365 091)

Murphy, S. & Grant, B. (1996). Portfolio approaches to assessment: Breakthrough or more of the same? In Edward M. White, William D. Lutz and Sandra Kamusikiri (Eds.), *Assessment of Writing: Politics, Policies, Practices*. New York: The Modern Language Association of America, 284-300.

Nunes, A. (2004). Portfolios in EFL classroom: disclosing an informed practice. *ELT Journal*, 58, 4, 327-335.

Oğuz, Ş. (2003). *State University Preparatory Class EFL Instructors' Attitudes Towards Assessment Methods Used at Their Institutions and Portfolios as a Method of Alternative Assessment*. Unpublished Master's thesis, Bilkent University: Ankara.

Paulson, F. L., Paulson, P. R., & Meyer, C. A. (1991). What makes a portfolio a portfolio? *Educational Leadership,* 48 (5), 60-63.

Paulson, F. L. & Paulson, R. R. (1994). *A guide for judging portfolios*. (ERIC Document Reproduction Service No. ED 377 210).

Raimes, A. (1991). Out of the woods: Emerging traditions in the teaching of writing. *TESOL Quarterly, 25*, 407-430.

Rodgers, T. S. (1989). After methods? What? In S. Avian (Ed.) *Language Testing Methodology*, pp. 1-15. Singapore: Regional English Language Centre.

Santos, M. (1997) Portfolio-based assessment. In Priscilla Butler (Ed.) *Issues in Alternative Assessment: The Japanese Perspective.* Kwansei Gaikun University, Nishinomiya (Japan) Language Center, 23-50.

Schneider, G. & Lenz, P. (2001). *European Language Portfolio: Guide for Developers.* http://culture2.coe.int/portfolio

Schwarz, J. A. & Collins, M. L. (1995). *Improving the Reliability of a Direct Writing Skills Assessment*. Paper presented at the Annual Meeting of the International Personnel Management Association Assessment Council Conference, New Orleans, LA. (ED 393880).

Shober, L. S. (1996). *A Portfolio Assessment Approach to Narrative Writing with the Cooperation of a Fourth Grade Target Group*. (ERIC Document Reproduction Service No. ED 395 318)

Song, B. & August, B. (2002). Using portfolios to assess the writing of ESL students: A powerful alternative? *Journal of Second Language Writing*, 11, 49-72.

Staehler, E. A. (1994). *Portfolio Assessment.* (ERIC Document Reproduction Service No. ED 393016)

Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press

Tedick, D. J. & Klee, C. A. (1998). Alternative Assessment in the Language Classroom. Washington, DC: Center for applied linguistics. (ERIC Document Reproduction Service No. ED 433720).

Tierney, R. J., Carter, M. A., & Desai, L. E. (1991). *Portfolio assessment in the reading-writing classroom.* Norwood, MA: Christopher-Gordon.

Trotman, W. (2004). Portfolio assessment: Advantages, drawbacks and implementation. *Testing Matters*, 13, 4, 62-65.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In Liz Hamp-Lyons (Ed.) *Assessing Second Language Writing* 111-125.

Webb, N. M. (1995). Group collaboration in assessment: Multiple objectives, processes and outcomes. *Educational Evaluation and Policy Analysis*, 17, 239-261.

Wiig, E. H. (2000). Authentic and other assessments of language disabilities: When is fair fair? *Reading & Writing Quarterly*, 16, 179-210.

Williams, J. D. (1998). *Preparing to Teach Writing Research, Theory, and Practice*. London: Lawrence Erlbaum Associates.

Wortham, S. C. (1998). Introduction. In Sue C. Wortham, Anna Barbour, Blanche Desjean-Perotta (Eds.) *Portfolio Assessment: A Handbook For Preschool and Elementary Educators*. (ERIC Document Reproduction Service No. ED 442 584.

APPENDIX A

INFORMED CONSENT FORM

Dear interviewee,

You have been asked to participate in a survey study which is intended to investigate the inter-rater reliability in writing portfolio assessment at Yıldız Technical University, School of Foreign Languages, Basic English Department. The study also aims at exploring the writing portfolio implementation in the institution.

In order to achieve the goals of the study, first you joined portfolio grading sessions and focus group discussions, which enabled us to use the current criteria and an analytic and holistic scale for portfolio grading. This interview will be the second phase of the study. You are going to be interviewed in order to have deeper insights of your perceptions of portfolio implementation and the 2 grading methods we have used in our grading sessions.

Your participation in the interview will bring valuable contribution to the findings of the study. Any information received will be kept confidential and your name will not be released. This study involves no risk to you.

I would like to thank you for your participation and cooperation.

Asuman Türkkorur

MA TEFL Program
Bilkent University

I have read and understood the information given above. I hereby agree to my participation in the study.

Name: ___________________________________

Signature: ________________________________

Date: ____________________________________

APPENDIX B

Analytic Scoring Scale

Content	
30-27	Excellent to very good: knowledgeable - substantive - thorough development of the thesis - relevant to assigned topic
26-22	*Good to average:* some knowledge of subject - adequate range - limited development of thesis - mostly relevant to topic, but mostly lacks detail
21-17	Fair to poor: limited knowledge of subject - little substance - inadequate development of topic
16-13	Very poor: does not show knowledge of subject - non-substantive - not pertinent - OR not enough to evaluate
Organization	
20-18	Excellent to very good: fluent expression - ideas clearly stated/supported - well-organized - logical sequencing - cohesive
17-14	Good to average: somewhat choppy - loosely organized but main ideas stand out - limited support - logical but incomplete sequencing
13-10	Fair to poor: non-fluent - ideas confused or disconnected - lacks logical sequencing and development
9-7	Very poor: does not communicate - no organization - OR not enough to evaluate
Vocabulary	
20-18	Excellent to very good: sophisticated range - effective word/idiom choice and usage - word from mastery - appropriate register
17-14	Good to average: adequate range - occasional errors of word/idiom form, choice, usage, but meaning not obscured
13-10	Fair to poor: limited range - frequent errors of word/idiom form, choice, usage - meaning confused or obscured
9-7	Very poor: essentially translation - little knowledge of English vocabulary, idioms, word form - OR not enough to evaluate
Language Use	
25-22	Excellent to very good: effective complex constructions - few errors of agreement, tense, number word order/function, articles, pronouns, prepositions
21-18	Good to average: effective but simple constructions - minor problems in complex constructions - several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions but meaning seldom obscured

17-11	Fair to poor: major problems in simple/complex constructions - frequent errors of negation, agreement, tense, number, word, order/function, articles, pronouns, prepositions and/or fragments - meaning confused or obscure
10-5	Very poor: virtually no master of sentence construction rules - dominated by errors, does not communicate, OR not enough to evaluate
Mechanics	
5	Excellent to very good: demonstrates mastery of conventions - few errors of spelling, punctuation, capitalization, paragraphing
4	Good to average: occasional errors of spelling, punctuation, capitalization, paragraphing but meaning not obscured
3	Fair to poor: frequent errors of spelling, punctuation, capitalization, paragraphing - poor handwriting - meaning confused or obscured
2	Very poor: no mastery of conventions - dominated by errors of spelling, punctuation, capitalization, paragraphing – handwriting, OR not enough to evaluate

APPENDIX C

Interview Questions:

1. What do you think about the portfolio implementation in your school? Is it just a folder kept by the student or an evidence of accurate representations of student work?
2. Do you have / need adequate training which will enable you to implement portfolio assessment in writing classrooms?
3. How do students gain understanding of (get feedback from) their writing? Teacher? Peers? Parents?
4. How much freedom of choice do you think students should have in deciding the portfolio content?
5. Do you think there is a match / consistency between teacher evaluations of student portfolios?
6. What do you think of the 2 grading methods we have used in our grading sessions?
    a. Do you feel these criteria match your subjective criteria for grading?
    b. Do you prefer to grade the portfolios the way you always have?
    c. Do you prefer standardized criteria?

APPENDIX D

Focus Group Discussion Questions

First Focus Group Discussion Questions

1. How is portfolio implementation working?
2. How do you feel about the 5% importance of portfolios on the overall grades of students? What percentage should portfolios have?
3. How do portfolios affect the improvement of students' writing abilities?
4. Do you think portfolios are a good indicator of students' writing abilities? How?
5. How do you think the portfolio process and grading affect student performance?
6. What are the most interesting topic items in the portfolios?
7. How could the portfolios be made more interesting?
8. Student portfolios can have various contents. Which one of these in the list on the board might you recommend in portfolios?
9. In the future which items might we include? How many?

Second Focus Group Discussion Questions

1. What goes into the grading in Portfolio Assessment?
2. What type of criteria did you use while grading the portfolios?
3. How would you weight the different sub-categories of the criteria?

4. What is your perspective on your own portfolio assessment criteria on the board?
5. What problems did you have that affect the reliability of the assessment?

Third Focus Group Discussion Questions

1. How does the new criteria work? / How do you feel about the new criteria?
2. Which sub-category affects you more in the whole portfolio grading?
3. What is the difference in your grading decisions between the first and second day?