T.C.
Mersin Üniversitesi
Sosyal Bilimler Enstitüsü
Yabancı Diller Eğitimi Ana Bilim Dalı

A STUDY ON ASSESSMENT TOOLS AND EVALUATION OF ESSAY WRITING SKILL IN FOREIGN LANGUAGE EDUCATION

Ulaş KAYAPINAR

DOKTORA TEZİ

Mersin, 2010

Mersin Üniversitesi, Sosyal Bilimler Enstitüsü Müdürlüğüne,

Ulaş KAYAPINAR tarafından hazırlanan "A Study on Assessment Tools and Evaluation of Essay Writing Skill in Foreign Language Education" başlıklı bu çalışma, jürimiz tarafından Yabancı Diller Eğitimi (İngiliz Dili Eğitimi) Anabilim Dalında DOKTORA TEZİ olarak kabul edilmiştir.

Başarılı	Başarısız		
☒	□	Başkan	Prof. Dr. Mustafa AKSAN (Danışman)
☒	□	Üye	Prof. Dr. Hatice SOFU
☒	□	Üye	Prof. Dr. Adnan ERKUŞ
☑	□	Üye	Doçent Dr. Adnan KAN
☒	□	Üye	Yrd. Doç. Dr. Şaziye YAMAN

Onay

Yukarıdaki imzaların, adı geçen öğretim elemanlarına ait olduklarını onaylarım.

.. / .. / ....

Prof. Dr. Mustafa AKSAN
Enstitü Müdürü

ACKNOWLEDGEMENTS

First of all, I’m so fortunate to have wonderful parents, mentors, colleagues, and friends. Each of them has been crucial to my success because, at one time or another, they have each taken a turn at challenging me, nurturing me, and teaching me. My deepest respect and indebtedness go to my advisors, Ayhan SEZER and Mustafa AKSAN, for their initial challenge and inspiration that made all the difference. I would also like to express my sincerest thanks to Adnan ERKUŞ and Adnan KAN for their knowledge, advice, and encouragement. Deserving special mention is Şaziye YAMAN who stood beside me throughout the long process. I would also like to acknowledge that Şaziye Yaman, Banu Tümay, Emsal Ateş Özdemir, Esin Ergene, Selvin Güven, Zeynep Yenigün, Suna Aksoy, Nurhayat Kavasoglu, Filiz Kalalı, Yıldız Turgut who devoted so much time to assessing the essays were very gracious participants in the study Mehmet Güngör. Finally, I owe a special debt to my wife because this study could never have been completed without her support and patience.

ABSTRACT

In this study, a valid and reliable scoring scale which can be used to assess essay writing skill was developed, and compared to the other assessing tools -general impression marking and checklist, used for essay writing skill considering scorer/rater reliability. In the development phase of the checklist and the scale, 103 faculty members took place. They evaluated the checklist and the scale criteria with performance indicators, and decided the acceptability levels of them in the scale. In the assessment phase, 10 scorers assessed 44 essays, written by ELT students, by using general impression marking, checklist, and scale. In order to determine the scorer reliability of each type of assessment, the relationship between the scorings was computed by using Pearson Product Moments correlation coefficient. Later, the results for different assessment tools were examined by converting to z scores by using Fischer’s z transformation. Employing a standardized open-ended question about each phase of the scoring process revealed the views and/or reactions of the scorers by their responses on using each scoring tool. Statistical analyses indicate that there is not a consistency between scorings done by using general impression marking. The scorings done by using checklist and scale include inconsistencies but the correlation coefficients are remarkably higher than the ones obtained from the scorings done by using general impression marking. There is a slight difference between the correlation coefficients obtained from the scorings done by using the checklist and the scorings done by using the scale in favor of the scale. However, no consistent and decisive scorings were done by the scorers using different tools in different time distances according to the results of Fischer’s z transformation. This may mean scorers assign different scores to the same essays when they use the same or different assessment tools in different time distances but scale use systematizes subjectivity by framing particular concerns of scorers while scoring.

Keywords: Writing, essay, assessment tool, scorer reliability.

ÖZET

Bu çalışmada essay yazma becerisini ölçmek için kullanılabilecek geçerli ve güvenilir bir puanlama ölçeği geliştirilmiş ve söz konusu ölçek essay yazma becerisini ölçmek için kullanılan diğer puanlama araçları –genel izlenim ve kontrol listesi- ile puanlayıcı güvenirliği açısından karşılaştırılmıştır. Kontrol listesi ve ölçeğin geliştirilme aşamasında üniversite öğretim elemanlarının görüşlerine başvurulmuştur. Yargıcılar Kontrol listesi ve ölçeği performans göstergelerine göre incelemiş ve uygunluk derecelerine karar vermişlerdir. Puanlama aşamasında ise 10 puanlayıcı İngilizce öğretmenliği bölümü öğrencilerinin yazdığı 44 essay ürününü genel izlenim, kontrol listesi ve ölçek kullanarak puanlamışlardır. Her puanlama türüne ilişkin puanlayıcı güvenirliğini belirlemek için puanlamalar arasındaki ilişki Pearson Momentler Çarpımı Korelasyon Katsayısı ile hesaplanmıştır. Farklı puanlama araçları için elde edilen sonuçlar Fischer'ın z dönüşümü kullanılarak z puanlarına dönüştürülerek incelenmiştir. Ayrıca, puanlayıcıların puanlamanın her bir aşamasına ilişkin görüşleri standard açık uçlu bir madde ile ortaya çıkarılmıştır. Analiz sonuçları genel izlenimle yapılan puanlamalar arasında bir tutarlılık olmadığını göstermiştir. Control listesi ve ölçekle yapılan puanlamalarda ise tutarsızlıklar olmakla birlikte genel izlenimle gerçekleştirilen puanlamalara göre çok daha yüksek korelasyon gösterdikleri görülmüştür. Control listesi ve ölçekle yapılan puanlamalar arasında ise küçük farklılıklar görülmüş ve sonuçların ölçekle puanlama lehine olduğu ortaya çıkmıştır. Buna rağmen, Fischer'ın z dönüşümü sonuçları farklı zamanlarda farklı puanlama araçları ile yapılan ölçmeler arasında tutarlı ve karalı puanlama yapılamadığını ortaya koymuştur. Bu durum, puanlayıcıların aynı essay ürünlerine farklı zamanlarda aynı ya da farklı ölçme araçlarıyla da puanlasalar farklı sonuçlar atadıklarını göstermektedir. Yine de, puanlamada ölçek kullanımının puanlayıcıları belli bir ölçüt ve puanlama çerçevesi içine alarak öznelliği sistematik hale getirdiği söylenebilir.

Anahtar Kelimeler: Yazma, essay, puanlama aracı, puanlayıcı güvenirliği.

TABLE OF CONTENTS

LIST OF TABLES

Page

INTRODUCTION

As the increase in the amount of intended behaviours and demands in academic life and the global world is rapid, it is clear that there is always an intensity of discussion and proposition about the issue of reliable testing and assessment. The store of knowledge grows at an enormous rate because there have been unavoidable changes and developments in the age of technology. The situation prompts educators to catch sophisticated moves forward because the amount of the objectives which are needed to be achieved for the necessity of continual development of the society increases.

In this respect, individuals and the society are in a permanent change and they should move together with the necessity mentioned in order to become fully-developed in light of education. In a climate of greater requirements for qualified individuals, testing each skill is uniquely difficult. More specifically, realiable and objective testing for language skills have been systematically developed but the debate on assessing writing skills still continues and it presents particular problems pertaining to validity and reliability. It is difficult to have control over writing and to evaluate writer's work objectively because measurement errors always exist in each measurement setting including the tester, the testee, the scoring method, and the scoring tool (Tekin, 2000). At this point, it is necessary to develop a scale for scoring as objectively as possible. However, the way it is done is one of the great difficulties of assessing writing even if there are vital decisions to be taken by assessing writing.

For this reason, the purpose of this study is primarily to develop a valid and reliable scoring scale which can be used to assess essay writing skill. The secondary purpose pertaining to the development of the particular scale, is to test the scorer/rater reliability of the scale by comparing it to other tools of assessing essay writing such as general impression marking and checklist. In order to introduce and test the particular measure for the assessment of essay writing skill, the following research questions are addressed in the present study.

Research Questions:

1. What are the characteristics of a scale used for assessing essay writing skill in English language teaching?
   a. What are the criteria of the scale used for a valid assessment?
   b. What are the weighted scores of the criteria in the scale?
2. How is the scorer reliability of the scale (ESAS)?
   a. How is the scorer reliability of the scale (ESAS) considering the total measurement results?
   b. How is the scorer reliability of the scale (ESAS) considering the results of sub-scorings?
3. How is the scorer reliability of general impression marking (GIM) considering the total measurement results?
4. How is the scorer reliability of the checklist (ECC) considering the total measurement results?

5. Is there any significant difference among the scorer reliability levels of the assessments?
6. What are the views of the scorers in assessment processes considering the assessment tools (GIM, ECC, and ESAS)?

Definitions

Rater/Scorer Reliability: The degree of consistency, agreement, or accuracy among inter-rater or intra-rater scorings of tests including the same or different items responded by different or same testees (Anastasi, 1982; Blok, 1985; van den Bergh & Eiting, 1989; Engelhard, 1996; Anastasi, 1997; Aiken, 2000; Congdon & McQueen, 2000; Johnson, Penny, & Gordon, 2001; Zhu, 2001; Burke & Dunlap, 2002; Atılgan, Kan, & Doğan, 2006; Eckes, 2008).

The Purpose and the Importance of the Study

The purpose of this study is to develop a valid and reliable scoring scale which can be used to assess essay writing skill, and to compare the particular scoring scale to general impression marking and checklist tools used in assessing essay writing skill according to scorer/rater reliability.

The importance of education is impossible to avoid if qualified individuals matter in the global world. For this reason, valid and reliable measurements and evaluation of educational products precede the determination of outcomes.

Especially, essay tests, which require writing on a specific topic in a pre-determined time, include items that may necessitate descriptions, explanations, comparison-contrasts, or judgments (Micheels, 1968: 240-241). The particular feature of essay items makes assessment difficult in an objective way. Research in this area shows that different or same raters independently assign different grades in different times (Karmel, 1970: 389; Marshall and Hales, 1972: 25; Gronlund, 1982: 72). This could be the evidence of unreliable assessments of this type of tests. Seen from this aspect, essay tests have an important role in both various degrees of education such as BA, MA, or PhD, and engagement of individuals. That the decisions should be made in valid and reliable methods is the building block of this study because decision makers may have personal or general judgments of testees out of purpose.

Moreover, a valid and reliable essay writing scale developed can provide unbiased measurements and assessments in the process of decision-making and then those decisions will be beyond any argument. For this reason, the study will contribute to foreign language education, measurement, and evaluation.

CHAPTER I

LITERATURE REVIEW

In this chapter, issues in language testing are discussed based on the general framework, planning, and its advantages. Later, the importance of the qualities of testing, validity and reliability, is explained. Next, writing evaluation and assessment are clarified in a general sense. Finally, assessing essay writing skill is discussed in order to emphasize the importance of scorer reliability and studies related to the topic.

I.1. Language Testing

The continual change of information network and existing knowledge in international competition necessitate that the educational system has to be developed and renewed with changing demands and conditions. Language teaching in this system has also experienced "many fluctuations and shifts over the years" (Celce-Murcia, 2001: 3) because it is essential to provide guidance and detail both at the classroom level and at processes of large-scale testing (Alderson, Clapham, & Wall, 1995; Bachman & Palmer, 1996; Weigle, 2002). Educational quality and standards at various stages of teaching and testing processes are also unavoidable (Prapphal, 2008: 140).

The increase in the amount of behavior indicators to be achieved in education has necessitated that the process which has been intended to provide second or foreign language learning has to be controlled and planned in a strict way. In order to realize the controlled and planned process, language teachers should ascertain the strengths and the weaknesses of the students. For different purposes of discovering the existence of some particular properties or for establishing different kinds of identities, testing is an accepted and essential part of all areas of life. As given great importance in social life and educational process, the reality of testing discipline also reflects the reality of language testing. Language testing cannot be considered apart from the teaching-learning process (Woodford, 1980). Information supplied from dependable measurements of language ability is certainly necessary to see the performance of the students and the reflections of the instruction. This might be the reason of the statement of Croft (1980: 473) as "the instructional activities, which are described as teaching, are testing itself". Testing can actually mean teaching because it is interrelated with all components of the language curriculum. It provides information for the examination and reexamination of each component of the curriculum. With relevance to the information supplied from testing, it is also called as the statements of how far needs are satisfied (Heaton, 1982). Besides, goals and objectives can be renewed or redesigned, teaching strategies and materials can be varied or differentiated, and the curriculum effectiveness and usefulness can be evaluated.

Language testing procedure includes two components which are measurement and assessment as each testing procedure does. The process that an amount of material learned or forgotten is measured by using tests and assigning a quantity is called measurement. The use of testing procedure and the activities of grading and classifying according to some specified criteria form assessment (Keeves, 1988). In other words, language teachers should determine the success levels of their students in acquiring the intended behaviour, and the success levels of the students can only be determined via the process of measurement and the assessment procedure including measurable objectives, decision-making, setting tasks, and scoring (Weigle, 2007).

Language teachers run testing process by using tests. Tests help them diagnose student strengths and weaknesses, assess students' progress, assist in evaluating students' achievement (Bachman, 1990: 3), and provide the control of entering or advancing to many important social roles.

If the students are given a statement of what they have achieved in a second or foreign language, then some kind of tests will be needed (Hughes, 1989). As Hughes states, the basic tools of the measurement and assessment process are tests. Tests are measuring instruments which are designed to measure the knowledge of the learners and their competence in language at particular times (Corder, 1973). Brown (2001) also states that a test is a method of measuring a person's ability or knowledge in a controlled area or activity of understanding.

Most of the testing plans are made according to a table of specifications to be sure that the items in the tests are equal in content and the content of the test reflects the content of the language curriculum (Bachman, 1990). Mcnamara (2000) gives emphasis on the instructions in the test and the structure of each part of the test. The format of response and the items that the students are supposed to be engaged are also important in the design of a test.

The significance of the test content could also be stated with a set of questions to be answered:

1. What kind of test is it to be?
2. What is its concrete purpose?
3. What behaviours are supposed to be measured?

(Hughes, 1989: 48)

In another way, the essential design of tests can be presented as a description of the test purpose, a description of the area of knowledge and types of tasks, a description of the intended test takers, a definition of the constructs to be measured, a plan for deciding the value of qualities or usefulness, and an inventory of necessary and available resources (Bachman & Palmer, 1996).

In stating the purposes for testing, Ur (1996: 34) focuses on it as a means to:

1. Give the teacher information about the present levels of the students
2. Give the students information about their own learning level and make them aware of their necessities
3. Assess for an external purpose of current teaching activities in the long term
4. Provide motivation for the students
5. Keep the concentration of the students on the subjects
6. Indicate the class has reached a position in learning
7. Lead the students to better results
8. Provide a useful review

9. Reveal a sense of achievement and progress in learning

By the purposes mentioned in the list above, students have useful information about their learning levels, strengths, and weaknesses. They also have motivation, feedback, and a sense of achievement and progress in learning. Besides, there are several advantages that teachers have by using tests because they use each test as a means to measure language proficiency, to reveal the achievement levels of the objectives, to identify the students' strengths and weaknesses, to show what they know and what they do not know, and to help in placing the students to appropriate teaching programme according to their language ability (Hughes, 1989). In addition, Brown (1995: 130) states the advantages of testing below:

1. To closely examine the perceptions of the students' needs
2. To concentrate on the remaining objectives to add new objectives designed to meet more advanced needs
3. To rethink the materials and teaching strategies to meet the newly perceived needs of the students
4. To have a great deal of information ready to be presented

In order to gain those advantages above, tests are designed in a way of planning to measure some kind of knowledge to supply some specific information. Besides, tests have powerful roles in many people's lives in the educational process and they are at significant transitional moments in education and beyond. Teachers rely on the information provided by the tests to make important decisions (McNamara, 2000). Moreover, the information supplied paves the way to make some kind of decisions on the students (Bachman & Palmer, 1996: 19). While making these decisions, teachers should be aware that the tests they use have the quality of measuring objectively. Expanded knowledge of the complexity of language test performance provide a basis for designing and developing language tests that are potentially more suitable for specific groups of testees and more useful for their intended purposes (Bachman, 1991). Determining the way the test measures the intended purpose consistently requires some kind of quality. The quality of a test can be defined by its validity and reliability.

I. 1.1. Validity

Curriculum is a broad concept. It includes all the teaching and learning activities in which students take part with the support of the school. This means the description of the subject area to be learnt, the ways of learning this subject area, and the ways that teachers give support in this process with the help of necessary materials, attributes, and methods of testing. At the end of the process, important decisions are made by the educators according to the measuring results. These results can indicate some deficiencies and this may cause consuming the time given, repeating the programme, severing the ties with the education process, higher expenses, depression, and losing motivation for the students. Besides, the students who are considered to have sufficient level to pass with wrong decisions mean higher expenses for universities or schools because of an individual cost and time-consuming irreparable drawbacks in the attempts of obtaining fully developed individuals. In the eyes of educators who keep the importance of the process in mind, the necessity of the quality of measuring tools which are used to make decisions can be seen effortlessly. Tests can be used to progress the curriculum by shaping the intentions and expectations of both the students and the teachers. Thus, they combine all parts of a curriculum such as cohesion, purpose, and control (Brown, 1995: 22).

There has been a continual development in language testing by a number of factors. These factors can be stated as the increasing number of the students which are involved in learning a second or foreign language initially. Secondly, the need for more accurate and comprehensive measures because of crucial importance of the tests increases in the educational process and real-life situations. Next, the emphasis on tests according to the usage of the language has changed. Fourthly, the scientific developments have led the way to measure the quality of the tests. Finally, the preparation of techniques for language test construction has developed continually (Croft, 1980). For this reason, validation in language testing is crucially important to ensure the fairness of the interpretations and the decisions on the performances of the students. It also involves the design, intentions, and logic of the test, and analysis on empirical evidence for the results and handling this evidence (McNamara, 2000).

Bearing in mind that a test is supposed to have consistent measurements performing its task, it should have the quality and attributes of measuring ability. The quality of a measurement tool or a test can primarily be determined by its validity. Connectively, an educator who develops a measurement tool should set up the conditions that the tool measures the intended property or behaviour in an exact way. In this way, it can realize the intended measurement without any error, thus the educator should examine the tool and determine its validity. Validity and reliability of a measurement tool should be stated before any application of design in the process.

Validity of testing is the most fundamental question which underlies the construction and use of assessments (Duran, 2008). The type of quality which reveals the degree that the test performs its task and reaches the target is called validity (Tekin, 2000). In another way, validity is a unitary concept which indicates the soundness of the decisions (Messick, 1992; Moss, Gerard, & Hanniford, 2006) and demonstrates that the test exactly measures the intended behaviour (Brown, 2001: 387). A high validity level is the indicator that the measuring tool has been developed for measuring a unique property aimed to be measured. The test must not measure anything else by chance or accidentally if it is to do its job properly (Corder, 1973). It is the indicator of the degree that the test measures specifically a unique property.

The tests which are valid help teachers understand who the test takers or their students are. They are the means of the evaluation process which the educator assures the objectives attained or not (Stern, 1983: 439). For this reason, the fairness and the clarity of the explanations and the decisions on the performances of the students should be provided without any doubt by validity in each measurement setting.

I. 1.2. Reliability

Reliability indicates the consistency and proximity of the test results obtained from the same individuals in different particular time distances (Anastasi, 1982: 102). It can be called as the adequacy of the measuring tool without any error even if it is applied at different times. If a group of students given the same test in different occasions has similar results, the test can be called as reliable (Brown, 1994: 253). In this way, it is necessary to examine whether the test performs its task by determining the true results or not. If it reaches the target, its reliability level can directly be known as sufficient.

In language teaching process, teachers try to understand who their students are in each language skill, especially at university level because mastery of a second or foreign language means communication in every aspect of life, and it is based on four basic language skills. However, it is possible that language tests do not have testing quality because of some primary factors affecting reliability.

Test length, expressions used in test instructions and items, homogeneity of items, and objectivity of scoring can be stated as primary factors affecting reliability (Baykul, 2000: 199). In the mean time, these factors affect validity. A lower reliability level is an indicator of a lower validity level but a higher reliability level may not be an indicator of a higher validity level because a test may have a sufficient reliability level but it may not measure the intended behaviours to be measured. To measure what is intended to be measured is the main property of validity.

One of the factors which can affect the reliability is the test length. It is assumed that the more item, as a representative indicator of the ability, to be measured, the more definite information as an indicator of that particular ability. Thus, if the number of items is increased, the reliability of the test spontaneously increases (Bachman, 1990: 220). It seems the test would have a lower reliability level if the reverse is the case.

Another factor is the expression used in test instructions and items. The expression of the instructions should be clear for the students. If not, each test taker may interpret the language used in the instruction in different ways and several different answers can occur beside the intended one (Turgut, 1997). Moreover, the items in the test should be clearly presented and easily understood because the reliability of the scores may be affected by ambiguous expressions of the items.

The next factor is the homogeneity of the items which means the items in the test should be homogeneous depending on the behaviours to be measured (Tekin, 1991: 70). The items in the test should be directly related to the content and the behaviours to be measured. A test including homogeneous items seems to be more reliable than another test including heterogeneous ones.

The last factor mentioned affecting reliability is the objectivity of the scoring. The consistency of scoring of a test by the same or different scorers in different times refers to the scoring or rater/scorer reliability. If a score obtained from a test has no change according to the scorer or according to the time elapsed, it has a high scoring reliability. Scoring reliability depends on the objectivity of the scoring; for this reason, objective tests have the highest scoring reliability (Tekin, 1991: 70) because scoring occurs after a scoring key or answer key is prepared. When scoring is made depending on the scoring key, the score does not change according to the scorer or the time elapsed.

Because writing tests include composing of thoughts or ideas according to a given topic or two, test length in writing tests cannot be used for a higher reliability. Test instructions may not contribute to the writings of the students in different times. There is no need to homogeneity of items because writing a paragraph or an essay includes implementing the same rules to different topics. Scoring reliability is hardly to be provided because scoring a test for writing skill depends on highly subjective assessments. On the other hand, if a detailed scoring scale is developed and scoring is realized depending on this scale writing tests may also become reliable.

Beside the primary factors affecting the reliability, some other random factors, which refer to the factors without prior planning, can be included such as testing environment, characteristics of the students, and changes in the attitudes of the students in time. These factors are difficult to control and they might not affect the reliability of a test with a valid importance but the primary factors can directly affect the reliability of a test. In measurement and assessment process of language skills, objective tests are intensively tried to be implemented because the factors mentioned above draw attention. For this reason, short answer and multiple choice items are usually used for testing language skills. However, it is not easy to say so for testing writing skills because a score obtained from a writing test can change according to the scorer and the time elapsed.

Validity and reliability are essential qualities of testing because they are the qualities that provide the major justification for using numbers or test scores as a basis for making inferences and decisions (Bachman & Palmer, 1996). These qualities of testing perform a vital task for language teaching since language tests are used as criterion measures of language learning abilities in second or foreign language teaching process. However, writing ability, especially the evaluation of writing, should be examined at this point.

I. 2. Writing Evaluation

Writing skill is the completing component of the mastery of a second or foreign language though it was seen as a supporting language skill until the early 1980s. It is a unifying instrument of thinking because it provides gaining control over thoughts, shaping perceptions, and recording information. Reid (1995) remarks that these features of writing were used in practicing courses such as grammar, reading comprehension, and vocabulary (Carter and Nunan, 2001), and observing the students' performance in the coursework. The situation supports the idea that the educators were partly unaware of the writing skill though it can be the indicator of high cognitive proficiency in the meantime.

A distinction is also proposed between the process of writing and the written product (Zamel, 1982; Chastain, 1988; Brown, 2001). The product approach of writing that the educators were concerned with in the early years of teaching writing refers to the final product of writing. Compositions which the students had written were supposed to include specific standards of the particular rhetorical style, to reflect the grammar in an accurate way, and to meet the expectations of the audience in conformity. Besides, the emphasis was on model compositions written by the students. In the process approach, writing instruction draws attention to five stages in the following (Scott & Vitale, 2003: 221):

1. Planning

2. Drafting

3. Revising

4. Editing

5. Publishing

Considering the list above, a writer first plans what to write, then makes a draft of it, later revises the draft, edits it, and finally publishes the writing. Simpson (2005) also suggests the stages as:

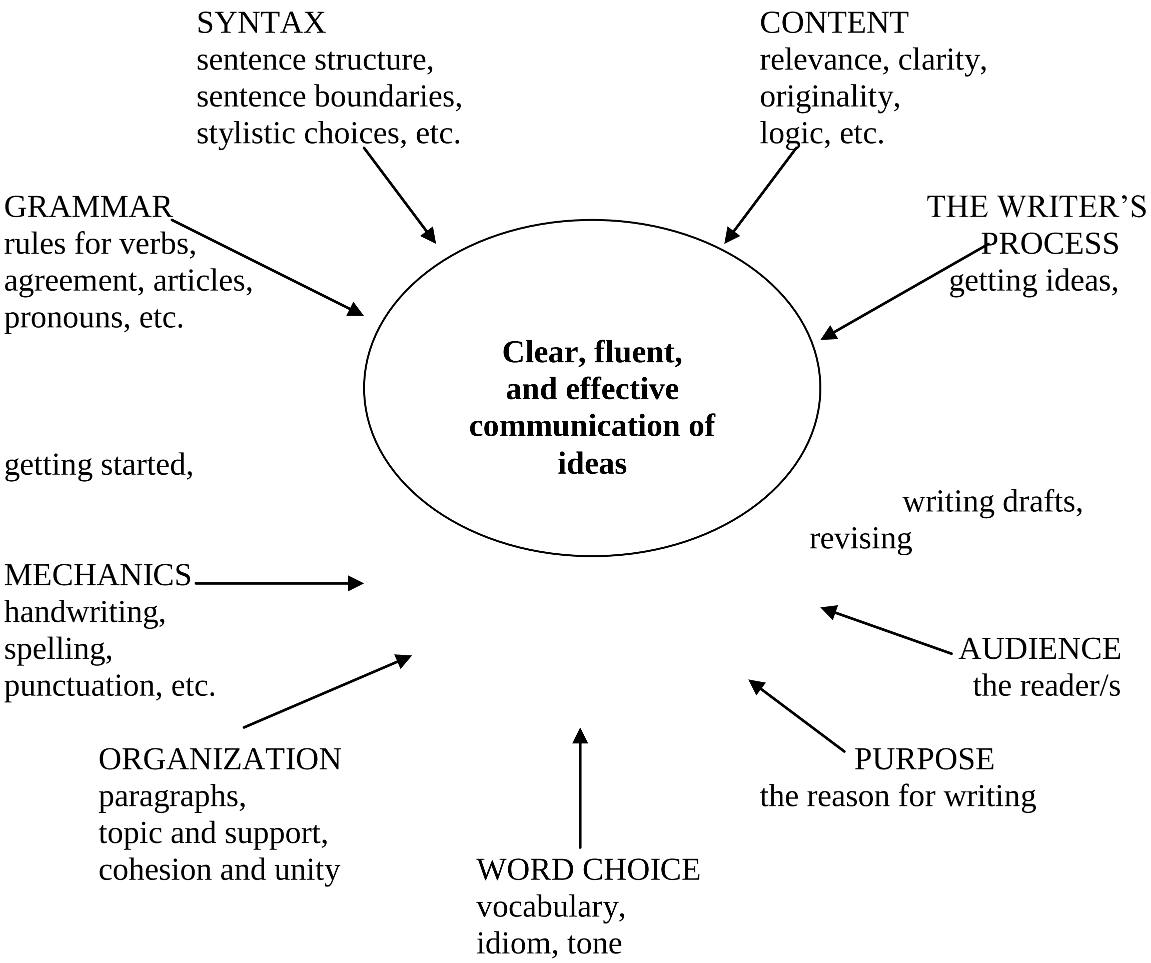
1. Prewriting (selecting a topic and planning what to say).
2. Writing (putting a draught version on paper).
3. Revising (making changes to improve the writing).
4. Evaluation (assessment of the written work).

In fact, both lists have the same stages in different words. Planning or prewriting is the first stage, putting a draft is the second stage, revising is the third one, and evaluating or editing is the next one. Scott and Vitale (2003) put the last stage as publishing which is missing in the former.

The process approach gives emphasis on the stages involved in creating a piece of work with a continuous and responsive interaction. Thus, the primary goal of the product approach is seen as an error-free coherent text. Besides, process approach allows for the fact that students will get closer to perfection by producing, discussing, and reworking successive drafts of a text. As the educators believe in educational development, they are also in charge of improving the written product. Therefore, the students have to be assisted in a way that will facilitate learning of the writing process by reaching the ability to produce the complete product. Since assisting the students and deciding their success levels accurately are considered important, evaluating the skill of writing which seems to be the determiner of the mastery of a second or foreign language at the highest level should be taken into consideration.

In order to cope successfully with the writing tasks, Celce-Murcia (2001: 211) states a set of specifications which includes the descriptions of task, content, and audience with format and linguistic cues, spelling and punctuation.

Raimes (1983: 6) also provides a diagram, which demonstrates the features that the writers deal with in the following:



The specifications of Celce-Murcia (2001) and Raimes' (1983) diagram give possible evidence for the categories included in the teaching process of writing skill. They can also give clues about the behaviours of essay writing to be measured.

Bachman (1990) mentions three assumptions required for the use of tests as a source of evaluation information. The first one is that educators undertake the necessity of information concerning educational outcomes for an effective education. The second one is that feedback provides appropriate changes to improve learning and teaching in the program. Lastly, the educational outcomes of the given program are measurable. The assumptions display the necessity of tests and measurable educational outcomes to provide information for an effective teaching learning process. In this respect, evaluating writing skill necessitates assessment by measuring the educational outcomes which is presented in the following part.

I. 2. 1. Assessing Writing Skill

Assessing writing is a difficult task (Bainer & Porter, 1992). As Bacha (2004) states, if a test is to provide meaningful information on which teachers can base their decisions, then many variables and concerns must be considered. For this reason, testing writing skill is seen as a problematic area. Miller and Crocker (1990) state that two groups exist in writing assessment: the supporters of direct methods and the supporters of indirect methods. Objective test items, especially in multiple-choice formats, are used in indirect methods which include the rules of word usage, spelling, grammar, punctuation, and sentence/paragraph structure (Breland *et al*, 1987). Weir (1990: 58-60) presents those indirect testing and direct testing. The first one is the "indirect testing" which includes measuring the elements of writing such as grammar, vocabulary, spelling, and punctuation by using objective tests. The second one is the "direct testing" of writing which includes measuring the student's ability to perform specific functional tasks such as essay writing by using subjective assessment. Moreover, original written compositions in response to a given topic are used in direct methods which include holistic, analytic, and primary trait scoring (Miller and Crocker, 1990). It is clear that subjectivity interferes in the assessment of writing. Therefore, it seems questionable that the success levels can be measured accurately with a subjective assessment. For this reason, teachers are always challenged by how to evaluate students' writing skills reliably and validly, so that the students will be better prepared for proficiency and achievement exams. Indeed, writing in the academic community can be called paramount; a student cannot be counted as successful in a second or foreign language without a certain level of academic writing proficiency. Then, the tests used in making decisions have to be both valid and reliable with a minimum level of subjectivity.

Writing tests have always been seen to be assessed subjectively because these tests are mostly assessed according to many different criteria or variables which are determined by the teacher of the particular students. If the teacher pays attention to different criteria while assessing the same tests, which belong to different students, this attitude increases the subjectivity of the assessment, inversely, it decreases the reliability level.

Three different scoring methods of assessing writing skills can be presented as general impression marking, holistic scoring, and analytic scoring:

1. General impression marking requires the teacher to read a sample of the papers to determine a standard and distinguish the students according to their achievement and according to the standards shaped in his/her mind. Two or more teachers giving an average mark based on the total impression of the paper (Hamp-Lyons, 1992) can also entail this scoring.

2. Holistic scoring, which dates back to the 1970s, refers to assessing the whole product according to prespecified levels and behaviour indicators (White, 1993; Yancey, 1999; Mertler, 2001). There is no individual trait scoring in holistic scoring, overall goals such as organization, content, and conventions are considered by the raters/scorers (Baldwin, 2004).

3. Analytic scoring refers to marking the papers according to some specific levels and criteria in a scoring scale, in another way, scorers assign scores to individual writing traits and they are able to score more sensitively (Weir, 1990; Mertler, 2001). Additionally, Hughes (1989) states that holistic scoring often refers to impressionistic scoring which involves assigning a single score to a piece of writing based on an overall impression. Otherwise, the analytic scoring requires a separate score for each of a number of aspects of the task, therefore; the analytic one is potentially more reliable and objective than holistic scoring (East, 2009).

Furthermore, Perkins (1983) suggests primary trait scoring. It evaluates features depending only on the particular audience and purpose of writing such as persuasiveness and clarity (Chastain, 1988). Thus, general impression marking and holistic scoring seem to be highly subjective because they depend on totally subjective impressions of the teacher, and analytic scoring can be used as a means of informing both the teacher and students of general and specific areas of high and low quality with a lower subjectivity level.

A test should have the property of scorability because the teacher can obtain the desired scores accurately and rapidly in the same manner for each exam paper (Gerberich, Greene, & Jorgensen, 1963). Moreover, a separate answer sheet or a scoring scale for the test itself can provide scorability of a test. Scoring scales draw attention to the categories to be measured that should be included in the criteria. Emerging from scoring need, different scoring scales and criterion categories have been tried to be developed for making essay tests more reliable with minimum subjectivity.

There have been many attempts to assess writing tests because assessing writing tests has been a major concern in marking objectively. Marking of these tests are always somewhat subjective at any rate, but marking scheme or scoring scales including descriptors can help make the marking consistent. Scoring scales are used to measure specific areas of language knowledge, which are determined from its operational definition by the developer. They include two parts:

1. Certain features of the language sample to be rated,
2. The definition of scale levels of the degree of mastery of the particular features (Bachman & Palmer, 1996: 212-213).

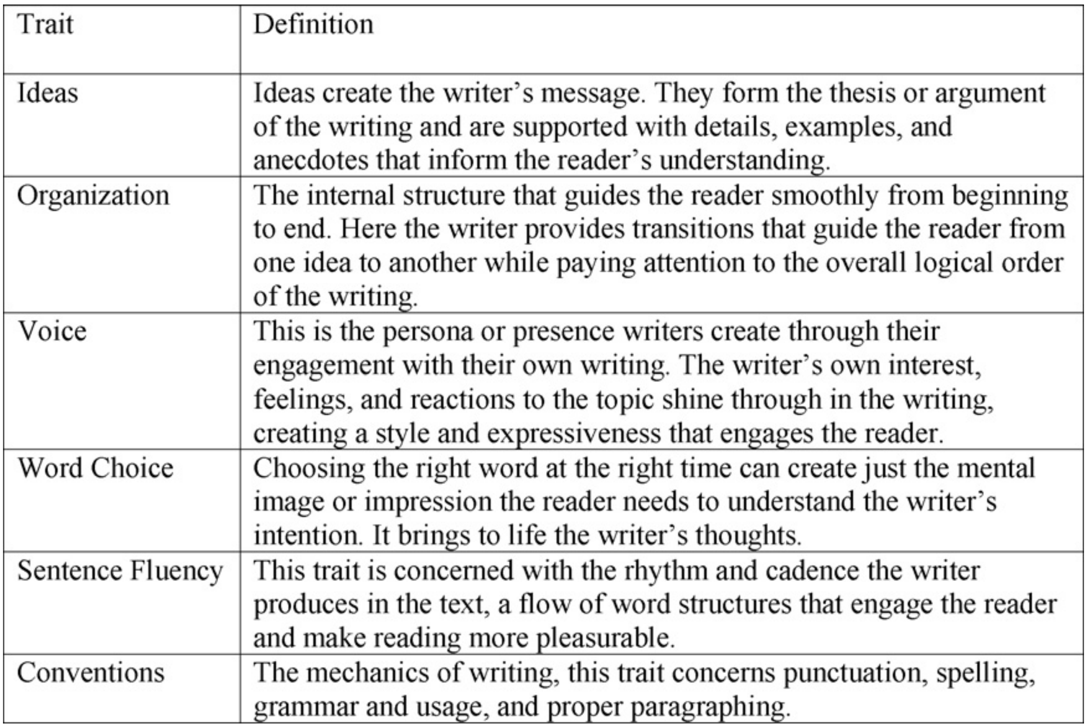
Throughout the history of writing scales, the Hillegas Composition Scale (1912) was one of the first attempts to measure an educational product for young learners' written expression. Thorndike (1914) extended the Hillegas scale to make it standardized. In 1920, M. H. Willing developed the Willing Scale for Measuring Written Composition which was made up of eight criteria of composition. More recently in 1950s, a group of publisher's staff members developed the STEP Essay Tests with the assistance of some advisors, readers, and test authors (Gerberich, Greene, & Jorgensen, 1963).

Bowen and Cali (2004) state that the North Carolina Writing Assessment scale, which was developed in the 1990s, included criteria such as main idea, support and elaboration, organization, and coherence. The scale was revised in 1999, and the criteria were extended as five features of effective writing; focus, organization, support and elaboration, style, and conventions.

Chapman (2004) also mentions that the rating system in Illinois Writing Program gives emphasis on focus, organization, support and elaboration, and conventions.

Spandel’s (2005) Six Trait Model is below seen as the basis for writing assessment program of Nebraska (Dempsey, PytlikZillig, & Bruning, 2009: 41).

Table 1. Spandel’s Six Trait Model



Trait	Definition
Ideas	Ideas create the writer’s message. They form the thesis or argument of the writing and are supported with details, examples, and anecdotes that inform the reader’s understanding.
Organization	The internal structure that guides the reader smoothly from beginning to end. Here the writer provides transitions that guide the reader from one idea to another while paying attention to the overall logical order of the writing.
Voice	This is the persona or presence writers create through their engagement with their own writing. The writer’s own interest, feelings, and reactions to the topic shine through in the writing, creating a style and expressiveness that engages the reader.
Word Choice	Choosing the right word at the right time can create just the mental image or impression the reader needs to understand the writer’s intention. It brings to life the writer’s thoughts.
Sentence Fluency	This trait is concerned with the rhythm and cadence the writer produces in the text, a flow of word structures that engage the reader and make reading more pleasurable.
Conventions	The mechanics of writing, this trait concerns punctuation, spelling, grammar and usage, and proper paragraphing.

As seen in the table, there are five traits called ideas, organization, voice, word choice, sentence fluency, and conventions. Each trait has a definition of the concern about the message of the writer. Moreover, Vaughan and Farr (2004) suggest that the scores received on an essay depend greatly on the bias of the teacher, and on other peripheral factors such as essay length, subject, and handwriting quality. In addition, they state the criteria as organization, completeness of thought, clarity of style, factual or illustrative support, correctness of grammar, punctuation, and mechanics. Brown (2001: 357) presents another list of criteria under the categories of content, organization, discourse, syntax, vocabulary, and mechanics for the evaluation of student writing. Brown also suggests a weighted scoring because of the need to assessing a grade or score to each paper as the following:

Content	: 0-24
Organization	: 0-20
Discourse	: 0-20
Syntax	: 0-12
Vocabulary	: 0-12
Mechanics	: 0-12
TOTAL	: 100

Bachman and Palmer (1996: 275-280) recommend another one which includes the knowledge of syntax, vocabulary, rhetorical organization, cohesion, and moderately formal register. Each category can be called as a sub-scale which is defined in terms of range and accuracy with five levels. Those levels give evidence about the mastery of the students as scoring "none (0), limited (1), moderate (2), extensive (3), and complete (4)". Madsen (1983: 121) presents another weighted scoring procedure by giving higher weight on grammar and usage together, and organization as the following:

Mechanics	: 20%
Vocabulary Choice	: 20%
Grammar & Usage	: 30%
Organization	: 30%

Some of the suggestions mentioned concern the main categories for the mastery of writing and some of them give percentage ratios which can be used as the determiners while assigning grades. Although each suggestion on scoring writings of the students includes the main categories to be measured, none of them can severe the ties with the subjectivity of scoring and provide the scoring reliability smoothly. Fisher, Brooks, and Lewis (2002) state fitness for purpose requirement is the core of all testing work, and direct writing assessments are subjective and thereby more prone to reliability issues. Given the potential limitations and suggestions for writing assessment, a more demanding aspect of reliability which is questioned in many respects is essay assessment.

I. 2. 2. Assessing Essay Writing Skill

Measuring and assessing essay writing skill in a reliable way has always been questioned in a broad sense because of the nature of essay tests. Essay tests and essay type items necessitate skills such as idea gathering, topic narrowing, paragraph writing, expressing a point of view, discussing a matter, and proving a thesis on any subject (Mackenzie, 2007: 6). This type of questions has been usually scored by three ways, as mentioned above, which are general impression marking, holistic scoring, and analytic scoring. General impression marking refers that the rater assesses the essays based on a single subjective standard for each and assigns a mark. This scoring type requires handling with the essay as a whole and a subjective judgment (Hamp-Lyons, 1992). Holistic scoring requires the teacher to score the overall product as a whole, with judging the predetermined level and component parts separately (Mertler, 2001; Nitko, 2001). In contrast, in analytic scoring, the teacher scores separate, individual predetermined criteria and levels of the product in the first step, then sums the individual scores to obtain a total score (Weir, 1990; Moskal, 2000; Nitko, 2001).

Vaughan (1991) states holistic scoring can be highly subjective and scores can vary in a significant way. Similarly, Hamp-Lyons (1991) suggests that holistic scoring may only be useful for some specific contexts. Huot (1990) also states that the levels of interrater reliability achieved with holistic scoring to be generally lower than that achieved with analytic scoring (Johnson, Penny, & Gordon, 2001). Veal and Hudson (1983), in their study, present evidence for higher reliability of analytic scoring than holistic scoring with a significant level. Breland's (1983) study also mentions that the use of holistic scoring may lead the way to great amount of differences in scores. Another study, by Johnson, McDaniel, and Willeke (2000), states that holistic scoring may result in great amount of differences in scores. In a study held by Gamaroff (2000), there is disagreement among the raters even in the categorization of errors. Some interpret an error as grammatical, others as lexical. Polio (1997) also examined consistency in scoring writing by using a holistic scale for 38 essays written by ESL undergraduate and graduate university students and the results indicated a low correlation between two raters using holistic measures (.44 and .53).

The fundamental concern questioned here is how to assess essays in a consistent and reliable way because there is always variation in the elements of writing preferred by the raters/scorers.

Some factors which may lead to those results are listed below:

1. Subjective decisions
2. Rater bias
3. Rating method
4. Personal beliefs about testees
5. Handwriting
6. Context position
7. Halo effect
8. Range restriction
9. Testee speed
10. Central tendency
11. Rater severity
12. Rater leniency
13. Testee characteristics
14. Essay length
15. Extraneous factors (Chase, 1968; Klein &Hart, 1968; Marshall & Powers, 1969; Hughes, Keeling & Tuck, 1980; Chase, 1983; Hughes, Keeling & Tuck, 1983; Hughes & Keeling, 1984; Blok, 1985; Wexley & Youtz, 1985; Sulsky & Balzer, 1988; Murphy & Balzer, 1989; Woehr & Huffcutt, 1994; Engelhard, 1994; Gyagenda & Engelhard, 1998a; Gyagenda & Engelhard, 1998b; Kan, 2005; Klein & Taub, 2005; Darus, 2006; Schaefer, 2008).

A form of resolving differences and inconsistencies in scoring papers, which has an improved interrater reliability, is recommended in various studies by using two or more raters in the scoring and computing the average of ratings (Coffman, 1971; Breland, 1983; Hieronymus, Hoover, Cantor, & Oberley, 1987; Herman, Aschbacher, & Winters, 1992; Cherry & Meyer, 1993; Kane, Crooks, & Cohen, 1999; Johnson, Penny, & Johnson, 2000; Johnson, Penny, & Gordon, 2001).

Another form is discussion presented as a method for scoring portfolios and for performing similar scorings when raters score independently (Cronbach, Lynn, Brennan, & Haertel, 1997; Johnson, Willeke, Bergman, & Steiner, 1997; Moss, Schutz, & Collins, 1997; Welch & Martinovich-Barhite, 1997; Clauser, Clyman, & Swanson, 1999). Here discussion technique is supported because of its power to create a dialogue, to promote critical reflection, and to constitute good teaching practice as a part of teachers' professional development (Johnson, Penny, & Gordon, 2001).

Maki (2004) suggests a training process until raters reach consensus to provide inter-rater reliability in the following steps:

- independent scoring
- discussion among raters to review responses
- discussion to reconcile differences
- repeating the process of independent scoring
- reviewing responses again
- discussion to reconcile differences

Another way to improve inter-rater reliability is to allow raters to augment their scores (Penny, Johnson, & Gordon, 2000). In this point of view, inter-rater reliability can be developed significantly over time with successive trials.

Think aloud protocols are also increasingly used as a source of data in a variety of research areas such as education and language teaching, memory operations, medical informatics, radiography nutrition, nursing, human computer studies, e-learning, and educational technology (Van Someren, Barnard, & Sandberg, 1994; De Larios, Murphy, & Manchon, 1999; Victori, 1999; Prime & Le Masurier, 2000; Whittington *et al*, 2000; Branch, 2001; Hartman, 2001; Cumming, Kantor, & Powers, 2002; Reicks *et al*, 2003; Jaspers *et al*, 2004*;* Cooke & Cuddihy, 2005; Schmitter-Edgecombe & Bales, 2005; Cotton & Gresty, 2006; Johnstone, Miller, & Thompson, 2006; Roberts & Fels, 2006; Klein, Schellings, Aarnoutse, & Van Leeuwe, 2006; Funkesson, Anbacken, & Ek, 2007; Göransson *et al*, 2007; Piacente-Cimini, & Williams, 2007; Banning, 2008; Hagen *et al*, 2008). They are defined as a useful source of data which can uncover the psychological mechanisms and knowledge structures underlying human problem-solving activities with respect to specific tasks by encouraging to "think-aloud"; to say what they are thinking and wondering at each moment (Yang, 2003: 96; Ericsson & Simon, 1984 in Cotton & Gresty, 2006).

In a study of formulation strategy in L2 composing, the students were instructed to verbalize all thinking while composing (De Larios, Murphy, & Manchon, 1999). In another study analyzing differences in the beliefs or metacognitive knowledge about writing, students are asked to think aloud while writing an argumentative essay in tape-recorded sessions (Victori, 1999). Teachers can also use these protocols by themselves for a lower subjectivity level while scoring essays or they can be defined as subjects whose internal responses are captured to reveal the reasons of differences or consistencies in scoring essays.

In this respect, general impression marking and holistic scoring can be called as subjective but analytic scoring can be called objective-like or systematically subjective because criteria indicators are scored subjectively. In both ways, there is a criterion or criteria set for assessment. Moreover, a scoring scale is needed in order to provide the scorability of essays consistently. Scoring scales are also called assessment scales, scoring rubrics, rating scales, and scoring guides (Aiken, 1996: 12; Bachman & Palmer, 1996: 212-213; Weigle, 2002: 118; Erkuş, 2006: 79; Knoch, 2009: 275). They include predetermined criteria, some kind of weighted scores or degrees, and partly performance indicators. Some of those can be seen in the following:

Table 2. Michigan Writing Assessment Scoring Guide

MICHIGAN WRITING ASSESSMENT SCORING GUIDE

English Composition Board: Criteria for Reading the Assessment

Ideas and Arguments	Rhetorical Features	Language Control
6 The essay deals with the issues centrally and fully. The position is clear, and strongly and substantially argued. The complexity of the issues is treated seriously and the viewpoints of other people are taken into account very well.	The essay has rhetorical control at the highest level, showing unity and subtle management. Ideas are balanced with support and the whole essay shows strong control of organization appropriate to the content. Textual elements are well connected through logical or linguistic transitions and there is no repetition or redundancy.	The essay has excellent language control with elegance of diction and style. Grammatical structures and vocabulary are well-chosen to express the ideas and to carry out the intentions.
5 The essay deals with the issues well. The position is clear and substantial arguments are presented. The complexity of the issues or other viewpoints on them have been taken into account.	The essay shows strong rhetorical control and is well managed. Ideas are generally balanced with support and the whole essay shows good control of organization appropriate to the content. Textual elements are generally well connected although there may be occasional lack of rhetorical fluency: redundancy, repetition, or a missing transition.	The essay has strong language control and reads smoothly. Grammatical structures and vocabulary are generally well-chosen to express the ideas and to carry out the intentions.
4 The essay talks about the issues but could be better focused or developed. The position is thoughtful but could be clearer or the arguments could have more substance. Repetition or inconsistency may occur occasionally. The writer has clearly tried to take the complexity of the issues or viewpoints on them into account.	The essay shows acceptable rhetorical control and is generally managed fairly well. Much of the time ideas are balanced with support, and the organization is appropriate to the content. There is evidence of planning and the parts of the essay are usually adequately connected, although there are some instances of lack of rhetorical fluency.	The essay has good language control although it lacks fluidity. The grammatical structures used and the vocabulary chosen are able to express the ideas and carry the meaning quite well; although readers notice occasional language errors.
3 The essay considers the issues but tends to rely on opinions or claims without the substance of evidence. The essay may be repetitive or inconsistent; the position needs to be clearer or the arguments need to be more convincing. If there is an attempt to account for the complexity of the issues or other viewpoints this is not fully controlled and only partly	The essay has uncertain rhetorical control and is generally not very well managed. The organization may be adequate to the content, but ideas are not always balanced with support. Failures of rhetorical fluency are noticeable although there seems to have been an attempt at planning and some transitions are successful.	The essay has language control which is acceptable but limited. Although the grammatical structures used and the vocabulary chosen express the ideas and carry the meaning adequately, readers are aware of language errors or limited choice of language forms.

successful.		
2 The essay talks generally about the topic but does not come to grips with ideas about it, raising superficial arguments or moving from one point to another without developing any fully. Other viewpoints are not given any serious attention.	The essay lacks rhetorical control most of the time, and the overall shape of the essay is hard to recognize. Ideas are generally not balanced with evidence, and the lack of an organizing principle is a problem. Transitions across and within sentences are attempted with only occasional success.	The essay has rather weak language control. Although the grammatical structures used and vocabulary chosen express the ideas and carry the meaning most of the time, readers are troubled by language errors or limited choice of language forms.
1 The essay does not develop or support an argument about the topic, although it may talk about the topic.	The essay demonstrates little rhetorical control. There is little evidence of planning or organization, and the parts of the essay are poorly connected.	The essay demonstrates little language control. Language errors and restricted choice of language forms are so noticeable that readers are seriously distracted by them.

(in Darus, 2006: 7-8)

The scoring guide in Table 2 has three criteria with six bands including performance indicators. The criteria are ideas and arguments, rhetorical features, and language control. The bands have performance indicators from 1 to 6 points explaining the ability of the writer in detail depending on the objectives. However, the noticeable differences among bands are made by using subjective expressions such as good, acceptable, clearly, well, and generally.

Table 3. Georgia Department of Education Holistic Rubric

GEORGIA DEPARTMENT OF EDUCATION **Stages of Writing Development Used in Holistic Rubric in the Scoring of Fifth-Grade Essays**	
Stages	**Description**
1: The emerging writer	Little or no topic development, organization, and/or detail Little awareness of audience or writing task Errors in surface features prevent the reader from understanding the writer's message
2: The developing writer	Topic beginning to be developed; response contains the beginning of an organizational plan Limited awareness of audience and/or task Simple word choice and sentence patterns Errors in surface features interfere with communication
3: The focusing writer	Topic clear even though development is incomplete; plan apparent although ideas are loosely organized Sense of audience and/or task Minimal variety of vocabulary and sentence patterns

	Errors in surface features interrupt the flow of communication
4: The experimenting writer	Topic clear and developed: development may be uneven; clear plan with beginning, middle, and end; beginning and/or ending may be clumsy Written for an audience Experiments with language and sentence patterns; word combinations and word choice may be novel Errors in surface features may interrupt the flow of communication
5: The engaging writer	Topic well developed; clear beginning, middle, and end; organization sustains the writer's purpose Engages the reader Effective use of varied language and sentence patterns Errors in surface features do not interfere with meaning
6: The extending writer	Topic fully elaborated with rich details; organization sustains writer's purpose and moves the reader through the piece Engages and sustains the reader's interest Creative and novel use of language and effective use of varied sentence patterns Errors in surface features do not interfere with meaning

(in Johnson, Penny, & Gordon, 2001)

The scale in Table 3 is also a holistic one which has six bands defining the writer as emerging, developing, focusing, experimenting, engaging, and extending. Each band has its own performance indicators without any criteria. However, the criteria are included in the bands such as surface features, use of language, audience, and organization.

Table 4. ESL Composition Profile

ESL COMPOSITION PROFILE		
CONTENT	30-27	EXCELLENT TO VERY GOOD: knowledgeable.substansive* thorough development of 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 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-substansive* not pertinent*OR not enough to evaluate
ORGANIZATION	20-18	EXCELLENT TO VERY GOOD: fluent expression * ideas clearly stated/supported * succinct * 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 form 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 forms * 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, run-ons, deletions* meaning confused or obscured
	10-5	VERY POOR: virtually no mastery 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 illegible* OR not enough to evaluate

(Jacobs *et al*, 1981 in Weigle, 2002: 116)

In Table 4, the scale presented has five main criteria which are content, organization, vocabulary, language use, and mechanics. Each criterion has its own bands which are called “very poor”, “fair to poor”, “good to average”, and “excellent to very good”. Each band has its own weighting range and holistic sub-criteria.

Table 5. Cambridge ESOL Main Suite and Other Exams

Cambridge ESOL main suite and other exams

Exam	Bands/levels	Main criteria for assessment
Certificate of Proficiency in English (CPE)	6 bands/levels Effect on reader Very positive Positive Achieves desired effect Negative Very negative Nil	Task realisation: content, organization, cohesion, range of structures, vocabulary, register and format, target reader; General impression: sophistication and range of language style, register, format organization and coherence topic development errors
Certificate in Advanced English (CAE)	6 bands/levels Effect on reader Very positive Positive Would achieve required effect Negative Very negative Nil	Task specific: Content; range; organization and cohesion; register; target reader; General impression: Task realisation: coverage, Resourcefulness Organization and cohesion Appropriacy of register Language: control, naturalness, range of vocabulary and structure, erros
First Certificate in English (FCE)	6 bands/levels Effect on reader Very positive Positive Would achieve required effect Negative Very negative Nil	Task specific: content; range; organization and cohesion; appropriacy of register and format; target reader; General impression: task realisation: full, good, reasonable, not adequate, not at all; coverage of points, relevance, omissions, original output Organization and links Control of language: range and accuracy Appropriacy or presentation and register
Preliminary English Test (PET) Key English Test (KET) Certificates in English Language Skills (CELS)	5 marks for test 5 for language 10, 5 and 5 marks for three tasks 6 bands/levels	Task coverage, elaboration, organization Language range, variety, complexity, errors Message communication, grammatical structure, vocabulary, spelling, punctuation Content points, length Format and register: Appropriacy Organization: clarity, intent Cohesion: complexity, variety of links Structure and vocabulary range: range± distortion Accuracy: ± impeding errors Paragraphing, spelling, punctuation
	9 bands/levels	Task fulfillment: requirements,

International English Language Testing (IELTS)	Effect on reader Expert Very good Good Competent Modest Limited Extremely limited Intermittent Non-user	exploitation, relevance, arguments, ideas, evidence: logic, development, point of view, support, clarity: coherence and cohesion Communicative quality: impact on reader, fluency, complexity, Vocabulary and sentence structure: range, appropriacy accuracy, error types

(in Hawkey & Barker, 2004: 130-131)

The holistic point of view on the scales is also seen in the criteria above. Each exam has its own criteria and bands or levels between zero and nine. The criteria are generally based on task realization, message communication, organization, format, register, structure and vocabulary range, and language use.

Table 6. TOEFL Writing Scoring Guide

TOEFL writing scoring guide

6 An essay at this level

- effectively addresses the writing task
- is well organized and well developed
- uses clearly appropriate details to support a thesis or illustrate ideas
- displays consistent facility in use of language
- demonstrates syntactic variety and appropriate word choice though it may have occasional errors

5 An essay at this level

- may address some parts of the task more effectively than others
- is generally well organized and developed
- uses details to support a thesis or illustrate an idea
- displays facility in the use of language
- demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors

4 An essay at this level

- addresses the writing topic adequately but may slight parts of the task
- is adequately organized and well developed
- uses some details to support a thesis or illustrate an idea

- demonstrates adequate but possibly inconsistent facility with syntax and usage
- may contain some errors that occasionally obscure meaning

3 An essay at this level may reveal one or more of the following weaknesses:

- inadequate organization or development
- inappropriate or insufficient details to support or illustrate generalizations
- a noticeably inappropriate choice of words or word forms
- an accumulation of errors in sentence structure and/or usage

2 An essay at this level is seriously flawed by one or more of the following weaknesses:

- serious disorganization or underdevelopment
- little or no detail, or irrelevant specifics
- serious and frequent errors in sentence structure or usage
- serious problems with focus

1 An essay at this level

- may be incoherent
- may be undeveloped
- may contain severe and persistent writing errors

0 A paper is rated 0 if it contains no response, merely copies the topic, is off-topic, is written in a foreign language, or consists of only keystroke characters.

(Weigle, 2002: 113)

There are also six bands in the scale above. Each band has different number of criteria. There are five criteria for Bands four, five, and six; band two and three have four criteria; and band one has three criteria. The criteria include task completion, organization, language use, syntactic variety, and word choice.

Table 7. Scoring Rubric for IELTS

Scoring rubric for IELTS

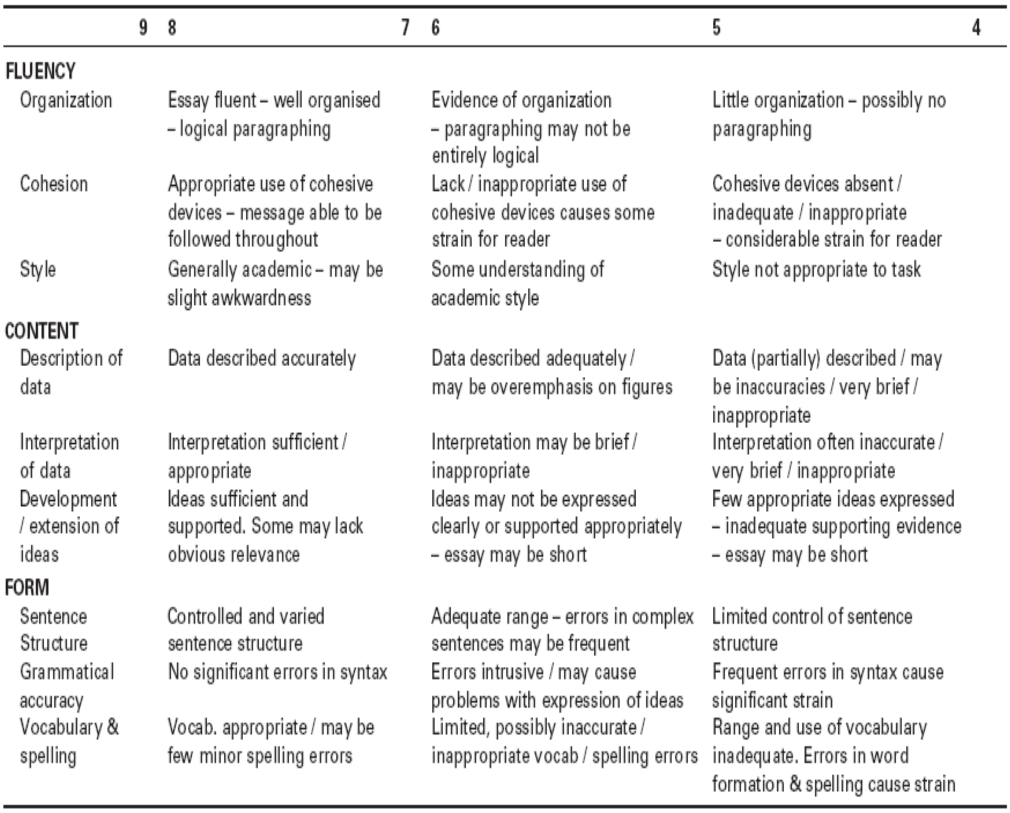
Has fully operational command of the language: appropriate, accurate and fluent with complete understanding.	**Expert user** **Band 9**
Has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriacies. Misunderstandings may occur in unfamiliar situations. Handles complex detailed argumentation well.	**Very good user** **Band 8**
Has operational command of the language, though with occasional inaccuracies, inappropriacies and misunderstandings in some situations. Generally handles complex language well and understands detailed reasoning.	**Good user** **Band 7**
Has generally effective command of the language despite some inaccuracies, inappropriacies and misunderstandings. Can use and understand fairly complex language, particularly in familiar situations.	**Competent user** **Band 6**
Has partial command of the language, coping with overall meaning in most situations, though is likely to make many mistakes. Should be able to handle basic communication in own field.	**Modest user** **Band 5**
Basic competence is limited to familiar situations. Has frequent problems in understanding and expression. Is not able to use complex language.	**Limited user** **Band 4**
Conveys and understands only general meaning in very familiar situations. Frequent breakdowns in communication occur.	**Extremely limited user** **Band 3**
No real communication is possible except for the most basic information using isolated words or short formulae in familiar situations and to meet immediate needs. Has great difficulty understanding spoken and written English.	**Intermittent user** **Band 2**
Essentially has not ability to use the language beyond possibly a few isolated words.	**Non user** **Band 1**
No assessable information provided.	**Did not attempt** **Band 0**

(IELTS, 2007: 4)

There are ten bands in the scale above which are defined as "did not attempt, non user, intermittent user, extremely limited, limited, modest, competent, good, very good, and expert. Each band has its own holistic criteria on the command of language and message completion.

Table 8. Diagnostic English Language Needs Assessment Rating Scale

Diagnostic English Language Needs Assessment Rating Scale



	9 8	7 6	5 4
FLUENCY			
Organization	Essay fluent – well organised – logical paragraphing	Evidence of organization – paragraphing may not be entirely logical	Little organization – possibly no paragraphing
Cohesion	Appropriate use of cohesive devices – message able to be followed throughout	Lack / inappropriate use of cohesive devices causes some strain for reader	Cohesive devices absent / inadequate / inappropriate – considerable strain for reader
Style	Generally academic – may be slight awkwardness	Some understanding of academic style	Style not appropriate to task
CONTENT			
Description of data	Data described accurately	Data described adequately / may be overemphasis on figures	Data (partially) described / may be inaccuracies / very brief / inappropriate
Interpretation of data	Interpretation sufficient / appropriate	Interpretation may be brief / inappropriate	Interpretation often inaccurate / very brief / inappropriate
Development / extension of ideas	Ideas sufficient and supported. Some may lack obvious relevance	Ideas may not be expressed clearly or supported appropriately – essay may be short	Few appropriate ideas expressed – inadequate supporting evidence – essay may be short
FORM			
Sentence Structure	Controlled and varied sentence structure	Adequate range – errors in complex sentences may be frequent	Limited control of sentence structure
Grammatical accuracy	No significant errors in syntax	Errors intrusive / may cause problems with expression of ideas	Frequent errors in syntax cause significant strain
Vocabulary & spelling	Vocab. appropriate / may be few minor spelling errors	Limited, possibly inaccurate / inappropriate vocab / spelling errors	Range and use of vocabulary inadequate. Errors in word formation & spelling cause strain

(in Knoch, 2009)

In Table 8, fluency, content, and form are the criteria which have sub-criteria such as organization, cohesion, style, data description, data interpretation, development of ideas, sentence structure, grammatical accuracy, vocabulary, and spelling. There are also bands from four to nine including performance indicators.

The criteria sets and bands given above underlie holistic scoring and checklist features. They also shed light on a subjective assessment and an intuitive approach. Fulcher (2003: 96) warns "the intuitive approach to scale development has led to a certain amount of vagueness and generality in the descriptors used to define bands". Relevantly, the key components of assessment should be defined and distinguished (Hawkey & Barker, 2004: 133); therefore, an attempt for construct definition, development of descriptors, and empirical analysis is needed in order to provide validity and reliability, and comparability. Huot (1990) states analytic type of scales or analytic scoring can be chosen as opposed to holistic scales or holistic scoring because:

- it is comparable with others
- it has proved to be more reliable than other types of scales
- it provides useful diagnostic information not found in other methods

(Sasaki & Hirose, 1999: 458)

McNamara (1996) notes that a scale represents the theoretical basis upon which the test is founded. This means it embodies the constructs or objectives stated to be measured implicitly or explicitly. In order to prove that analytic type of scales are more reliable and accurate to use in essay scoring, each study should be primarily held by utilizing the decisions made by experts for determining the theoretical basis of the scale to be used. In this respect, the next chapter is presented to clarify the method of the study.

CHAPTER II

METHOD

This chapter includes the type of the research, participants of the study, data gathering instruments, procedure, data analysis, assumptions, and limitations.

II.1. Type of the Research

The primary purpose of this study is to develop a checklist and a rating scale for essay writing assessment. The secondary purpose is to examine the scorer reliability of essay writing skill by using those assessment tools mentioned and general impression marking. The study gives emphasis on whether the use of a scoring scale provides an objective and reliable scoring of essay writing skill or not. However, this study is a fundamental research because the results of this study cannot be generalized for the population of the scope, it is only for providing new findings and results contributing the current ones.

II.2. Participants

Because the study is a fundamental research, there is not any sample representing a population. Instead, three main study groups were employed:

1. Judges include 103 faculty members of ELT departments from different (20) universities in Turkey. They evaluated the checklist (ECC, see App. 1) and the scale (ESAS, see App. 2) criteria and performance indicators, and decided the acceptability levels of them in the scale.

2. Scorers (10 people) who assessed the essays were ELT experts (MAs and PhDs) and experienced teachers of writing skill.

3. ELT students (44 people) who responded the essay test were from Mersin University Advanced Reading and Writing class.

II.3. Data Gathering Instruments

The basic data gathering instruments are presented below:

1. ***44 essays***. Written for the final exam of Advanced Reading and Writing class in Mersin University ELT Department in order to achieve the objective below:

"By means of the awareness of essay types, essay writers will analyze, synthesize and evaluate information and therefore, in their compositions, react to prompts. Essay writers will also be able to analyze and produce different types of essays (e.g. comparison and contrast, classification, process analysis, cause-and-effect analysis, and argumentative) that are unified, coherent and organized." The essay prompt, which was determined by the teachers of the particular class, is the same for all students as given in the following:

Please write an essay about the topic below:

"University students should be free to choose their own courses."

2. ***Essay Criteria Checklist* (ECC)**. First of all, a criteria list was developed by the researcher through a review of relevant literature (Raimes, 1983; Greenberg and Rath, 1985; Norton, 1990; Celce-Murcia, 2001; Johnson, Penny, & Gordon, 2001; Jacobs *et al*. 1981 in Weigle, 2002; Weigle, 2002; Bowen and Cali, 2004; Hawkey & Barker, 2004; Darus, 2006; IELTS, 2007; Dempsey, PytlikZillig, & Bruning, 2009; Knoch, 2009;). In the next step, 103 faculty members of ELT departments from different (20) universities examined the checklist considering the expressions used and the consistency between the objectives of essay writing skill and the items included. Later, 2 experts of measurement and evaluation examined the checklist considering the technical features. Finally, an essay criteria checklist was developed by the researcher using expert judgments after a last check.

3. ***Essay Assessment Scale* (ESAS)**. First, 103 faculty members of ELT departments from different (20) universities around the country examined the scale considering the expressions used and the consistency between the objectives of essay writing skill and the performance indicators included. Next, 2 experts of measurement and evaluation examined the scale considering technical features. Finally, a Likert type scale covering five performance levels (0-1-2-3-4) was developed by expert judgments. In fact, there is no limit for performance levels, however, five performance levels was chosen because of easiness and usefulness for the observable behaviour (Kan, 2007).

4. The measurement results of general impression marking

5. The measurement results by using ECC

6. The measurement results by using ESAS

7. The written views of the scorers provided by the responses to a standardized open-ended question about the scoring process.

II.4. Procedure

The procedure of the study includes two phases given below:

1. ***The preparation of the material to be scored***. The essays were asked by the teacher of Advanced Reading and Writing class in Mersin University ELT Department to be written by 44 students in their class time for final exam. As soon as the application finished, the essay responses were collected by the teacher, and the researcher took them in order to get them ready for scorings. The essays written by the students were rewritten by the researcher by using a word processor program and saved. This procedure was realized in order to avoid the bias emerging from the effect of handwriting. All other issues relating to the writings of the students kept the same.

2. ***Scorings.*** The design of the scorings are presented in the table below.

Table 9. Scoring Design

	Scorers				
Scorings	**1**	**2**	**3**	**…**	**10**
GIM	44 ESSAYS	44 ESSAYS	44 ESSAYS	…	44 ESSAYS
10 WEEK-BREAK					
GIM	44 ESSAYS	44 ESSAYS	44 ESSAYS	…	44 ESSAYS
10 WEEK-BREAK					
ECC	44 ESSAYS	44 ESSAYS	44 ESSAYS	…	44 ESSAYS
10 WEEK-BREAK					
ECC	44 ESSAYS	44 ESSAYS	44 ESSAYS	…	44 ESSAYS
10 WEEK-BREAK					
ESAS	44 ESSAYS	44 ESSAYS	44 ESSAYS	…	44 ESSAYS
10 WEEK-BREAK					
ESAS	44 ESSAYS	44 ESSAYS	44 ESSAYS	…	44 ESSAYS

As seen in Table 9, there are ten scorers who participated the study. Those scorers scored 44 essays by using general impression marking, essay criteria checklist, and essay assessment scale in different time distances from one scoring to another. These scorings were employed after a 10 week-break in order to remove the memory effect of the previous scoring and an educational session for each scoring held independently by the researcher. In order to provide objectivity, students' names were not written on the essays and a code number was assigned to each essay. Moreover, these code numbers were shuffled for each scoring.

II.5. Data Analysis

In the study, expert judgments and the percentage values were used in order to determine the criteria used in the scale. 103 experts from 20 different universities assessed the appropriate criteria for the use of the scale.

In order to determine the scorer reliability of each type of assessment, the relationship between the scorings was computed by using Pearson Product Moments correlation coefficient. The results were examined by using Fischer’s z transformation. This procedure led the way to set the scorer reliability coefficients provided by different scoring techniques in order.

Employing a standardized open-ended question about each phase of the scoring process reveals the views and/or reactions of the scorers by their responses on using each scoring tool. The qualitative data here were analyzed line by line and memos were written (Strauss & Corbin, 1998; Glesne, 1999). Categories or labels were reviewed and recurring themes, core consistencies and meanings were identified by using pattern codes. Then those explanatory pattern codes were identified as smaller sets, themes or constructs with content analysis (Miles & Hubermas, 1994; Patton, 2002). The process is as follows:

- Underlying key terms in the responses
- Restating key phrases
- Coding key terms
- Pattern coding
- Constructing themes
- Incorporating themes into an explanatory framework

II.6. Assumptions

The assumptions of the study are given below:

1. The judges presented their own views for the development of the checklist with devoutness.
2. The judges presented their own views for the development of the scale with devoutness.
3. The scorers scored the essays with devoutness in a serious manner.
4. Each scoring is independent from the others.
5. The scores assigned to the essays by using different scoring tools are independent from each other.

II.7. Limitations

The study is limited with:

1. The judgments of 103 ELT faculty members in 20 different universities in Turkey,
2. 44 essay products of 44 ELT students from Mersin University,
3. The scoring results of 10 scorers by using GIM, ECC, and ESAS, and
4. The views of the scorers.

The method given in this chapter leads the way to the findings which are discussed in the next chapter by following the research questions.

CHAPTER III

FINDINGS AND DISCUSSION

In this chapter, the findings and discussions are presented by following the research questions.

III.1. Findings and Discussion of Research Question 1

"What are the characteristics of a scale used for assessing essay writing skill in English language teaching?"

The main criteria of the scale for a valid assessment, which were determined by 103 experts, are presented below:

1. Organization
2. Language Use
3. Vocabulary
4. Mechanics
5. Ideas/Content

These main criteria have some sub-criteria as follows:

1. ORGANISATION

1.1. Introduction (Introductory Sentences, Thesis Statement)

1.2. Body Paragraphs (Topic Sentence, Supporting Sentences)

1.3. Conclusion

2. LANGUAGE USE

2.1. Word Order

2.2. Pattern Variety

2.3. Verb Form

2.4. Tenses

2.5. Articles

2.6. Pronouns

2.7. Prepositions

3. VOCABULARY

3.1. Word Choice

3.2. Word Variety

3.3. Parts of Speech

4. MECHANICS

4.1. Punctuation

4.2. Capitalization

4.3. Paragraphing

4.4. Indentation

5. IDEAS/CONTENT

5.1. Title

5.2. Development

5.3. Unity

5.4. Transitional Signals

The criteria indicators are determined as follows:

1. ORGANISATION

1.1. Introduction

1.1.1. Introductory Sentences : Effective introductory sentences

1.1.2. Thesis Statement: Appropriate thesis statement (thesis and central idea)

1.2. Body Paragraphs

1.2.1. Topic Sentence : Appropriate topic sentence (possibly implied) supporting the thesis and the central idea

1.2.2. Supporting Sentences : Appropriate sentences supporting the topic (possibly major and minor)

1.3. Conclusion : Appropriate conclusion related to thesis

2. LANGUAGE USE

2.1. Word Order: Correct word order

2.2. Pattern Variety: Using different patterns

2.3. Verb Form : Using verb forms correctly

2.4. Tenses: Using tenses appropriately

2.5. Articles: Using articles correctly

2.6. Pronouns: Using pronouns correctly

2.7. Prepositions: Using prepositions correctly (verb + preposition, adjective + preposition)

3. VOCABULARY

3.1. Word Choice: Selecting the appropriate words

3.2. Word Variety: Having a rich vocabulary

3.3. Parts of Speech : Using the correct parts of speech

4. MECHANICS

4.1. Punctuation: Using punctuation marks correctly

4.2. Capitalization: Using cases (lower/upper) correctly

4.3. Paragraphing: Correct paragraph formatting

4.4. Indentation: Using margins correctly and consistently

5. IDEAS/CONTENT

5.1. Title: Appropriate title

5.2. Development: Appropriate development

5.3. Unity : Unity

5.4. Transitional Signals: Using appropriate transitional signals

The weighted scores of the main criteria are in the following:

Table 10. The Weighted Scores of Organization Criterion

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	4,00	9	8,7	8,7	8,7
	5,00	94	91,3	91,3	100,0
	Total	103	100,0	100,0	

As seen in the table, ninety-four teachers ‘totally agree’ on the criterion. Moreover, nine of them state that they ‘agree’ on the criterion.

Table 11. The Weighted Scores of Language Use Criterion

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3,00	2	1,9	1,9	1,9
	4,00	13	12,6	12,6	14,6
	5,00	88	85,4	85,4	100,0
	Total	103	100,0	100,0	

Table 11 indicates that eighty-eight teachers totally agree on the criterion. Only 2 of them are undecided.

Table 12. The Weighted Scores of Vocabulary Criterion

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3,00	2	1,9	1,9	1,9
	4,00	24	23,3	23,3	25,2
	5,00	77	74,8	74,8	100,0
	Total	103	100,0	100,0	

It is clear in the table that two teachers are undecided and one hundred and one of them agree on the criterion.

Table 13. The Weighted Scores of Mechanics Criterion

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2,00	3	2,9	2,9	2,9
	3,00	10	9,7	9,7	12,6
	4,00	20	19,4	19,4	32,0
	5,00	70	68,0	68,0	100,0
	Total	103	100,0	100,0	

Seventy of the teachers agree totally on mechanics criterion, whereas, three of them state it is not an appropriate criterion for essay scoring.

Table 14. The Weighted Scores of Ideas/Content Criterion

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3,00	3	2,9	2,9	2,9
	4,00	5	4,9	4,9	7,8
	5,00	95	92,2	92,2	100,0
	Total	103	100,0	100,0	

In Table 14, ninety-five teachers state that ideas/content criterion is a must for essay scoring. Only three of them are undecided and they do not have an idea for it.

The weighted scores of the sub-criteria are presented in Table 15:

Table 15. The Weighted Scores of Sub-Criteria

	SUB-CRITERIA	FREQUENCY				
		1	2	3	4	5
ORGANIZATION	A.1.Introduction	1	0	0	4	98
	A.1.1. Introductory Sentences	0	1	3	17	82
	A.1.2. Thesis Statement	0	0	2	11	90
	A.2. Body Paragraphs	1	0	1	11	90
	A.2.1. Topic Sentence	0	0	1	8	94
	A.2.2. Supporting Sentences	1	1	2	19	80
	A.3. Conclusion	0	0	4	9	90
LANGUAGE USE	B.1. Word Order	0	0	5	27	71
	B.2. Pattern Variety	0	4	20	23	56
	B.3. Verb Form	0	1	10	27	65
	B.4. Tenses	0	1	8	24	70
	B.5. Articles	2	8	22	26	45
	B.6. Pronouns	1	5	13	22	62
	B.7. Prepositions	1	6	19	28	49
VOCABULARY	C.1. Word Choice	0	1	5	35	62
	C.2. Word Variety	0	5	13	33	52
	C.3. Parts of speech	2	4	16	21	60
MECHANICS	D.1. Punctuation	1	8	17	24	53
	D.2. Capitalization	1	3	24	22	53
	D.3. Paragraphing	0	5	6	23	69
	D.4. Indentation	0	10	19	27	47
IDEAS/CONTENT	E.1. Title	2	5	10	17	69
	E.2. Development	0	0	3	23	77
	E.3. Unity	0	1	4	11	87
	E.4. Transitional Signals	1	3	5	24	70

As seen in the table, the judges agree on each criterion at the level of seventy percent at least. These results refer to the weightings below for the main criteria with equally scored sub-criteria on a 4-point Likert type scale out of 100 points:

Organization	: 28
Language Use	: 28
Mechanics	: 16
Ideas/Content	: 16
Vocabulary	: 12

III.2. Findings and Discussion of Research Question 2

"How is the scorer reliability of the scale?"

The results (out of 100 points) of the scorings done by the scorers by using the scale developed are presented below:

Table 16. ESAS Scoring Results.

	S*1.1**	**S1.2**	**S2.1**	**S2.2**	**S3.1**	**S3.2**	**S4.1**	**S4.2**	**S5.1**	**S5.2**	**S6.1**	**S6.2**	**S7.1**	**S7.2**	**S8.1**	**S8.2**	**S9.1**	**S9.2**	**S10.1**	**S10.2**
1.	84	84	66	60	92	91	61	92	72	73	82	78	87	87	51	58	80	83	70	72
2.	75	75	78	69	80	73	76	80	76	77	64	63	84	91	84	79	86	85	89	90
3.	76	65	51	57	50	91	51	82	53	56	46	45	69	70	39	48	56	47	58	60
4.	51	40	68	68	79	82	60	60	67	68	74	56	80	82	81	90	82	80	80	80
5.	65	55	80	80	79	79	82	82	83	80	75	47	88	87	92	92	86	82	78	81
6.	64	65	54	57	59	65	57	67	56	59	47	61	86	89	45	46	62	67	65	70
7.	70	70	74	74	66	71	53	69	67	68	52	67	83	89	58	49	87	93	72	81
8.	57	57	66	69	55	58	62	71	64	70	66	42	85	80	50	55	61	67	67	66
9.	68	70	64	69	43	54	57	74	69	63	61	56	69	71	47	52	61	62	67	70
10.	52	55	80	72	78	82	61	79	77	76	85	49	92	91	73	72	78	76	63	60
11.	78	75	79	74	95	93	55	88	66	72	52	51	86	92	40	40	83	83	71	75
12.	70	70	61	50	69	71	52	72	68	70	54	52	68	66	43	48	46	58	59	60
13.	65	67	71	69	83	85	54	84	63	65	64	53	71	82	43	46	68	90	86	85
14.	85	84	67	68	84	78	54	97	59	53	71	46	70	69	69	60	64	85	65	74
15.	68	59	64	65	67	66	58	69	53	59	46	53	82	90	32	41	72	73	58	77
16.	70	75	67	64	85	90	69	76	66	69	50	58	70	66	40	45	63	68	55	70
17.	68	60	72	73	85	74	78	91	72	71	56	40	55	53	76	81	82	90	60	60
18.	65	50	58	55	80	66	68	74	67	67	68	43	70	73	44	40	72	86	58	70
19.	84	65	73	70	88	83	51	79	65	63	88	68	84	58	71	67	68	75	67	66
20.	80	70	58	55	77	86	51	85	70	70	54	53	75	74	51	45	78	68	54	54
21.	55	72	67	51	82	72	61	80	70	60	55	47	86	80	52	46	74	81	64	70
22.	76	60	82	70	97	85	68	80	77	87	80	81	89	91	54	59	85	88	79	81

	S1.1	S1.2	S2.1	S2.2	S3.1	S3.2	S4.1	S4.2	S5.1	S5.2	S6.1	S6.2	S7.1	S7.2	S8.1	S8.2	S9.1	S9.2	S10.1	S10.2
23.	50	52	75	63	91	91	63	71	75	74	77	77	91	91	59	59	84	90	68	70
24.	75	75	67	63	92	90	64	78	68	69	61	62	85	90	51	48	56	48	62	71
25.	75	65	48	69	62	69	57	75	62	68	55	52	76	79	38	39	40	46	46	52
26.	68	55	64	65	88	78	58	74	72	71	51	50	85	89	59	51	59	63	76	80
27.	55	45	62	62	82	82	54	79	68	64	72	52	79	73	57	44	71	71	67	74
28.	67	63	64	52	76	80	61	85	67	67	73	76	86	85	60	61	65	64	56	62
29.	68	50	60	57	86	80	56	86	75	73	60	48	85	89	40	43	76	73	55	53
30.	57	60	56	53	64	73	60	79	62	60	53	49	72	92	37	44	63	66	54	65
31.	65	70	71	71	65	69	63	75	72	75	74	74	89	92	53	55	77	73	84	92
32.	68	70	62	62	71	84	57	81	77	78	61	64	88	92	47	47	72	75	58	52
33.	67	66	54	65	75	85	54	73	58	67	52	44	63	69	37	40	62	66	51	54
34.	70	70	63	61	84	87	59	70	59	63	76	39	53	90	34	43	63	52	56	57
35.	81	90	59	55	97	91	71	76	64	67	75	67	91	86	67	61	74	59	71	82
36.	70	70	69	58	71	85	73	73	68	68	82	71	56	65	53	49	65	70	70	75
37.	85	87	71	65	81	78	87	69	75	74	53	50	72	89	60	65	62	72	48	53
38.	85	65	47	63	51	74	60	69	87	79	48	49	50	89	58	58	60	70	48	59
39.	76	80	85	80	84	89	70	69	66	61	86	74	90	80	71	74	84	83	79	82
40.	84	82	62	65	52	77	61	91	57	57	52	51	64	75	46	48	82	80	61	72
41.	85	85	65	54	67	79	51	73	64	67	58	50	63	91	45	42	83	86	68	70
42.	88	90	69	69	57	82	67	86	64	56	62	69	71	75	59	61	74	74	63	59
43.	76	80	72	68	80	83	69	83	64	68	76	70	88	91	74	61	87	89	78	85
44.	79	85	84	73	67	89	54	76	72	72	53	48	52	54	52	52	59	61	56	59

*Scorer, **Scoring

As seen in Table 16, same or similar results are sometimes seen between two scorings. In order to see the consensus between the two scorings done by the same scorers, intra-scorer consensus is presented below in Table 17.

Table 17. Intra-scorer consensus between scorings by using ESAS.

Difference	Scorer 1		Scorer 2		Scorer 3		Scorer 4		Scorer 5		Scorer 6		Scorer 7		Scorer 8		Scorer 9		Scorer 10	
	f	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**f**	**%**
0	9	21	5	11	3	7	3	7	6	14	2	5	2	5	6	14	3	7	3	7
±1-5	18	41	21	48	18	41	3	7	28	64	18	41	24	55	24	55	23	53	24	55
±6-10	8	18	4	9	11	25	6	14	10	23	7	16	8	18	12	28	12	28	10	23
±11-15	4	9	14	32	7	16	7	16	0	0	6	14	2	5	2	5	4	9	6	14
±15-more	5	11	0	0	5	11	25	57	0	0	11	25	8	18	0	0	2	5	1	2
TOTAL	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100

Table 17 shows that Scorer 1 scored 27 essays out of 44 with ±0-5 points difference out of 100. This means 62% of the essays have similar results in two scorings done by using ESAS. In the scorings of Scorer 2, the number of the essays scored with ±0-5 points difference is 26, and the percentage is 59%. Scorer 3 scored 21 essays with ±0-5 points difference, which means 48%. Scorer 4 is the one who has the smallest amount of consistency. The scorer scored only 6 essays with ±0-5 points difference, which refers to 14%. In the scorings of Scorer 5, the number of the essays scored with ±0-5 points difference is 34, which is quite high (78%) when compared to others. The results of Scorer 6 show that 20 essays were scored with ±0-5 points difference out of 100. Scorer 7 scored only 2 essays the same but there are 26 essays scored with a ±0-5-point difference out of 100. Scorings of Scorer 8 show 30 essays have ±0-5 points difference out of 100, which refers to 69%. In the scorings done by Scorer 9, the number of essays with ±0-5 points difference out of 100 is 26. Finally, Scorer 10 scored 27 essays with ±0-5 points difference with a percentage of 62.

For a better understanding of the scorer reliability of the scale, it is necessary to examine the correlation coefficients between the two scorings done by using ESAS.

The correlation coefficients computed, by using Pearson Product-Moment Correlation, between the first and the second scorings and they are presented below in Table 18:

Table 18. Correlations across ESAS scorings.

Scorer	Correlation coefficient
1	. 757**
2	. 641**
3	. 585**
4	. 021
5	. 825**
6	. 680**
7	.545**
8	. 916**
9	. 811**
10	. 884**

** Correlation is significant at the 0.01 level.

The results indicate that the correlation coefficients between the scores scorers assigned to the essays seem to be high and significant at the 0.01 level (no less than .545) except the one which was done by Scorer 4 (.021). These results refer that 9 scorers scored the essays in a consistent way. Moreover, 7 of the correlation coefficients are around .70. This is a high level of positive correlation which is seen meaningful and which may mean that there is a high consistency between those scorings (Kline, 1986).

When the results are compared to the ones in Table 17, Scorer 4 is the one who has the smallest amount of intra-scorer consistency, correspondingly, the one whose results have the lowest and the only insignificant correlation coefficient. The highest correlation coefficient belongs to Scorer 8 (.916) whose scores correspond to each other. This refers to similar results for two scorings done in different time distances. Moreover, Scorer 8 is the one who scored 42 essays out of 44 with ±10 points difference out of 100 (intra-scorer consensus=95%). This is the best result among the scorers, however, the differences among the correlation coefficients, even the ones within a 10-point difference in total scores, of the same essays scored in different times indicate there is always a source of variation in scorings done by ESAS.

Seven scorers submitted the sub-scoring results and the correlation coefficients which belong to sub-scorings were computed as in the following:

Table 19. Correlations across ESAS sub-scorings.

Criteria		**Scorer**					
	2	**4**	**5**	**6**	**7**	**9**	**10**
Organization	.734**	.159	.410**	.316*	.154	.431**	.218
Language Use	.436**	.200	.349*	.339*	.186	.271	.044
Vocabulary	.425**	.164	.325*	.414**	.154	.378*	.206
Mechanics	.559**	.143	.298*	.438**	.150	.192	.141
Ideas/Content	.328*	.409**	.363*	.391**	.324*	.231	.322*

*Correlation is significant at the 0.05 level
**Correlation is significant at the 0.01 level

Table 19 indicates some information about the significance of the correlation coefficients. This means there are some positive relationships between the scorings of sub-criteria except the scorings done by Scorer 4, 7, and 10. It seems that 4 correlation coefficients are significant in scorings of organization criterion. However, the correlation coefficients are low compared to the others for the total results. This result can show that the results of those scorings are hardly similar to each other. The highest correlation coefficient (.734) between the first and the second scorings of the criterion seems to be obtained by the scoring of Scorer 2. This may indicate that there is a high relationship between the scores assigned for each scoring. The scorings done by Scorer 4 (.159), 7 (.154), and 10 (.218) seem to have no relationship between two scorings. This means that it is not advantageous for those scorers to use a scale in scoring essays. In the meantime, there may be some other sources of variation played role in scorings. The same results can be seen in the other scorings of the other criteria considering the same scorers (4, 7, and 10) except the ideas/content criterion. For this criterion, all correlation coefficients seem to be significant in acceptable levels (0.05 and 0.01) with only one exception. It was computed for the scorings done by Scorer 9 (.231). Additionally, because the correlation coefficient is not significant for the scorings of language use and mechanics criteria, Scorer 9 (with values of .271 and .192) joins Scorer 4 (.200 and .143), Scorer 7 (.186 and .150), and Scorer 10 (.044 and .141) in the scorings of these criteria.

III.3. Findings and Discussion of Research Question 3

“ How is the scorer reliability of measurement results by using general impression marking (GIM)?”

First of all, the results (out of 100 points) of the two scorings of the scorers are presented below:

Table 20. GIM Scoring Results.

	S*1.1**	**S1.2**	**S2.1**	**S2.2**	**S3.1**	**S3.2**	**S4.1**	**S4.2**	**S5.1**	**S5.2**	**S6.1**	**S6.2**	**S7.1**	**S7.2**	**S8.1**	**S8.2**	**S9.1**	**S9.2**	**S10.1**	**S10.2**
1.	60	85	40	42	65	80	60	50	88	82	65	60	70	85	50	65	85	75	68	65
2.	70	85	55	65	85	90	90	85	73	84	60	55	65	70	80	70	90	85	63	80
3.	55	80	29	30	40	65	50	50	64	68	60	50	55	90	40	40	75	70	60	64
4.	60	70	75	76	75	75	60	50	66	65	78	73	54	60	55	85	90	90	70	75
5.	65	65	78	78	80	95	95	80	88	92	80	75	50	65	95	90	90	90	75	79
6.	55	85	44	33	65	75	60	45	64	68	65	58	58	75	55	40	60	60	57	62
7.	60	65	69	50	87	95	65	65	60	78	82	78	58	72	60	90	75	80	70	77
8.	60	75	62	50	70	75	60	65	71	64	75	70	60	70	50	40	65	45	54	60
9.	50	65	47	50	50	70	65	60	72	68	70	65	65	20	45	50	55	50	55	65
10.	70	65	79	82	75	85	65	75	76	80	83	78	95	98	65	50	60	70	80	87
11.	70	80	75	62	75	80	65	70	83	80	72	70	97	76	60	55	85	90	73	80
12.	80	80	78	59	65	85	65	85	74	84	87	85	60	10	50	50	65	40	53	50
13.	75	75	80	74	65	90	65	70	78	88	72	66	52	7	75	70	70	30	80	79
14.	55	80	65	38	66	70	60	70	84	90	60	60	45	58	60	35	70	40	70	70
15.	55	50	57	52	68	90	65	70	84	74	85	80	55	98	80	60	80	80	79	80
16.	60	65	48	43	55	85	65	75	82	64	64	80	80	75	50	50	75	60	80	75
17.	60	65	45	40	80	80	60	80	82	78	78	76	48	15	45	60	70	50	64	69
18.	70	70	56	35	77	80	80	70	82	75	66	70	45	57	55	55	70	40	56	64
19.	65	55	50	38	80	85	75	70	86	62	70	75	78	100	65	65	85	85	64	69
20.	80	65	48	50	65	75	70	70	84	76	60	58	35	5	65	65	50	20	50	57
21.	60	55	63	65	35	65	45	45	77	66	58	53	40	15	30	50	55	20	59	67
22.	75	55	70	67	70	95	80	85	80	73	85	80	70	99	50	55	70	25	84	89

	S1.1	S1.2	S2.1	S2.2	S3.1	S3.2	S4.1	S4.2	S5.1	S5.2	S6.1	S6.2	S7.1	S7.2	S8.1	S8.2	S9.1	S9.2	S10.1	S10.2
23.	60	75	34	65	65	80	60	60	86	65	62	57	30	30	50	60	55	55	45	50
24.	60	50	60	74	70	80	70	70	67	55	68	60	90	99	65	65	70	70	55	66
25.	75	70	58	58	60	75	55	85	68	68	62	60	75	18	80	60	50	25	55	64
26.	80	70	40	68	75	85	75	60	78	86	73	70	50	78	65	80	75	70	76	84
27.	65	65	46	53	68	85	90	70	78	74	80	75	60	20	60	75	80	70	69	75
28.	70	55	87	52	78	90	85	80	78	84	85	78	97	50	60	70	80	75	79	77
29.	50	65	52	49	80	80	65	45	82	76	68	60	43	85	70	75	80	75	45	55
30.	75	65	64	28	70	80	65	60	84	60	62	57	68	93	50	80	80	75	60	60
31.	85	55	80	74	85	85	55	60	94	80	75	68	98	100	55	70	85	80	81	80
32.	80	60	63	56	85	95	70	65	88	74	68	72	99	100	60	90	90	85	78	75
33.	75	65	60	55	80	95	75	65	92	70	60	65	96	98	60	60	80	65	60	60
34.	75	65	28	45	70	60	50	55	79	64	65	58	78	72	55	75	35	30	51	50
35.	85	65	78	55	90	75	65	70	95	75	72	66	95	80	45	95	60	50	89	90
36.	80	80	68	40	80	65	65	80	92	73	60	65	78	50	55	65	65	45	67	60
37.	70	70	49	47	80	80	70	40	90	62	65	68	83	78	75	60	60	65	69	75
38.	65	85	53	41	85	95	65	70	94	85	71	74	94	95	70	80	75	80	60	70
39.	65	85	50	41	75	80	60	55	80	76	57	64	92	100	60	60	70	60	55	60
40.	80	75	77	75	79	75	60	80	94	90	84	77	40	50	80	95	80	55	75	81
41.	50	80	38	45	80	80	60	65	88	84	75	68	99	100	50	55	75	65	50	61
42.	45	42	30	35	65	85	50	55	79	55	60	55	65	5	30	40	30	25	45	52
43.	70	85	32	67	67	75	50	60	78	48	63	58	42	5	30	50	35	30	45	55
44.	55	70	72	75	68	90	60	80	75	72	80	75	97	95	45	85	75	85	69	55

*Scorer, **Scoring

In Table 20, it is very difficult to see same or similar results between two scorings. In order to see the consensus between the two scorings done by the same scorers, examining intra-scorer consensus below may be more helpful.

Table 21. Intra-scorer consensus between scorings by using GIM.

Difference	Scorer 1		Scorer 2		Scorer 3		Scorer 4		Scorer 5		Scorer 6		Scorer 7		Scorer 8		Scorer 9		Scorer 10	
	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%
0	7	16	2	5	6	14	6	14	1	2	1	2	1	2	9	21	7	15	3	7
±1-5	9	21	17	38	8	18	18	41	13	30	30	68	10	23	7	15	1	6	18	41
±6-10	8	18	7	15	9	21	8	18	13	30	12	27	6	14	7	15	7	15	14	32
±11-15	9	21	6	14	8	18	4	9	6	14	0	0	6	14	9	21	2	5	2	5
±15-more	11	25	12	27	13	30	8	18	11	25	1	2	21	47	12	27	12	27	7	15
TOTAL	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100

Table 21 shows that Scorer 6 scored 31 essays out of 44 with ±0-5 points difference out of 100. This is the highest value among the others referring that 70% of the essays have similar results in two scorings done by using GIM.

The scorings of Scorer 9 has the lowest percentage of consensus which is 18% with a ±0-5-point difference. The frequency is also 1 for zero difference, and 7 for ±1-5-point difference. Other scorers' consensus between two scorings by using GIM has a frequency range between 11 and 21 points. Table 21 also indicates that the percentages of the scores which are the same in two scorings have a range between 2 and 21. This means that the frequencies range between 1 and 9 out of 44 essays. Scorer 5, 6, and 7 have only one score which is the same for two scorings. However, Scorer 8 scored 9 essays the same.

For a better understanding of the scorer reliability of general impression marking, it is necessary to examine the correlation coefficients between the two scorings done by using GIM. The correlation coefficients computed, by using Pearson Product- Moment Correlation, between the first and the second scorings and they are presented below in Table 22:

Table 22. Correlations across GIM scorings.

Scorer	Correlation coefficient
1	.042
2	.510**
3	.477**
4	.279
5	.450**
6	.835**
7	.584**
8	.412**
9	.790**
10	.880**

** Correlation is significant at the 0.01 level

The correlation coefficients, seen in Table 22, range between .042 and .880. Among the ten coefficients, 2 of them, which belong to the scorers 1 and 4, are not significant. The other correlation coefficients seem significant. This may mean that those scorers gave similar scores to the essays in both scorings. However, only 3 of them are above .70 which refer to a high and meaningful correlation coefficient and relatively a high consistency.

III.4. Findings and Discussion of the Research Question 4

"How is the scorer reliability of measurement results by using the checklist (ECC)?"

In the first step, it is useful to see the results (out of 100 points) of the second two scorings of the scorers. The results are presented below:

Table 23: ECC Scoring Results.

	S1.1	S1.2	S2.1	S2.2	S3.1	S3.2	S4.1	S4.2	S5.1	S5.2	S6.1	S6.2	S7.1	S7.2	S8.1	S8.2	S9.1	S9.2	S10.1	S10.2
1.	84	82	70	67	80	70	68	67	78	73	65	68	87	30	49	51	81	83	74	77
2.	76	77	75	79	90	91	91	81	82	78	70	74	85	78	63	87	85	88	85	89
3.	64	81	54	50	30	35	65	65	60	58	54	59	70	65	50	47	60	56	75	71
4.	60	52	81	86	88	90	83	83	64	60	72	68	86	95	90	88	88	83	84	90
5.	80	67	71	74	95	96	87	82	88	85	76	70	80	99	95	93	76	80	88	80
6.	60	65	49	48	30	75	77	73	62	62	58	50	88	70	50	53	62	63	74	70
7.	72	73	66	69	60	80	65	71	62	68	70	60	65	70	49	50	70	80	89	79
8.	60	58	55	52	40	70	76	68	66	64	76	69	45	90	53	54	76	70	70	77
9.	64	68	48	44	48	65	69	82	64	62	74	70	30	10	55	53	67	62	74	79
10.	75	54	86	89	80	85	76	78	80	77	85	77	90	99	61	63	72	70	94	88
11.	80	56	52	55	76	65	71	90	77	74	75	70	78	90	42	45	84	86	89	80
12.	72	78	58	53	80	85	93	81	78	75	87	80	30	83	53	50	59	63	64	57
13.	78	51	53	50	72	90	91	92	87	88	67	64	20	50	54	56	49	42	90	79
14.	48	75	45	39	60	61	66	69	90	90	59	63	30	15	59	57	51	60	84	70
15.	63	74	62	58	85	79	79	71	76	77	89	84	88	93	77	75	80	81	90	84
16.	60	68	58	49	61	78	72	71	60	60	63	60	68	85	46	45	75	80	84	77
17.	75	59	49	55	84	71	82	74	73	70	70	74	35	30	48	45	76	73	76	70
18.	60	68	62	56	76	87	76	78	68	66	65	65	30	15	50	55	67	61	72	65
19.	69	69	68	65	80	73	78	70	58	50	74	70	70	80	64	68	87	91	80	84
20.	52	59	44	41	68	65	79	88	70	74	65	63	20	30	50	48	50	53	68	60
21.	56	65	51	43	56	73	71	84	57	60	66	62	20	10	46	50	45	52	76	64
22.	76	69	50	53	93	80	66	70	75	72	85	88	80	90	43	40	44	50	92	87

	S1.1	**S1.2**	**S2.1**	**S2.2**	**S3.1**	**S3.2**	**S4.1**	**S4.2**	**S5.1**	**S5.2**	**S6.1**	**S6.2**	**S7.1**	**S7.2**	**S8.1**	**S8.2**	**S9.1**	**S9.2**	**S10.1**	**S10.2**
23.	48	68	66	60	56	65	64	74	59	57	75	70	15	15	60	65	66	60	60	60
24.	64	70	60	55	76	84	58	73	50	50	68	62	95	95	39	40	67	67	74	69
25.	68	83	74	72	68	60	65	80	72	70	69	65	60	10	50	48	37	36	78	70
26.	60	74	53	50	88	60	74	78	78	74	84	82	15	10	76	78	73	76	88	81
27.	56	88	42	40	64	75	79	83	70	70	72	66	70	15	55	59	75	78	84	81
28.	64	87	66	69	90	85	78	86	76	75	83	80	60	30	70	75	73	71	88	79
29.	60	77	50	44	64	80	75	81	70	65	61	55	60	55	51	55	64	72	69	68
30.	48	84	44	41	90	84	73	74	62	60	64	60	75	55	57	63	80	81	72	75
31.	60	85	63	69	88	92	70	80	78	80	73	70	96	97	57	63	81	85	90	88
32.	64	89	79	82	80	94	78	78	80	75	75	70	100	98	62	66	85	92	88	83
33.	68	77	54	56	88	88	72	69	75	72	69	67	78	95	55	58	71	74	74	75
34.	68	80	47	43	80	65	58	62	68	70	63	58	60	75	44	43	58	63	60	60
35.	86	85	41	40	90	87	80	87	80	75	65	68	65	97	56	55	65	70	94	82
36.	72	80	41	45	64	82	86	71	75	76	70	65	75	70	50	53	86	74	74	78
37.	88	71	67	64	84	71	70	60	70	72	74	70	90	90	56	60	80	76	84	82
38.	84	68	68	70	80	66	74	58	80	78	74	76	90	99	48	46	73	82	80	79
39.	68	84	55	55	74	64	74	59	72	70	60	54	50	95	46	43	76	71	70	74
40.	76	67	56	61	72	90	76	64	86	85	73	71	50	50	49	55	79	80	90	86
41.	80	71	43	39	80	77	72	61	85	85	77	79	98	98	53	57	76	75	72	80
42.	80	68	41	45	74	70	62	56	58	55	71	70	40	20	42	40	45	50	61	58
43.	72	69	57	60	81	65	71	58	55	60	81	80	20	10	55	51	56	48	65	60
44.	80	82	59	62	82	77	78	68	75	78	83	86	100	85	56	60	90	91	70	80

In Table 23, it is not usually easy to see same or similar results between two scorings. In order to see the consensus between the two scorings done by the same scorers, examining intra-scorer consensus below may be more helpful.

The frequencies and percentages are presented below:

Table 24. Intra-scorer consensus between scorings by using ECC.

Difference	Scorer 1		Scorer 2		Scorer 3		Scorer 4		Scorer 5		Scorer 6		Scorer 7		Scorer 8		Scorer 9		Scorer 10	
	f	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**f**	**%**
0	1	2	1	2	1	2	3	7	6	14	1	2	5	11	0	0	1	2	2	5
±1-5	9	21	35	80	14	32	14	32	36	82	33	75	9	21	40	91	30	69	21	48
±6-10	12	27	8	18	7	16	14	32	2	5	10	23	9	21	3	7	12	27	17	39
±11-15	6	14	0	0	10	23	11	25	0	0	0	5	11	25	0	0	1	2	4	9
±15-more	14	32	0	0	12	27	2	5	0	0	0	16	36	82	1	2	0	0	0	0
TOTAL	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100

Table 24 shows that Scorer 5 scored 42 essays out of 44 with a ±0-5-point difference out of 100, although there are 6 essays scored with a zero difference. This is the highest value among the others referring that 96% of the essays have closer results to each other in two scorings done by using ECC.

The scorings of Scorer 1 has the lowest percentage of consensus which is 23% with a ±0-5-point difference. The frequency is also 1 for zero difference, and 9 for ±1-5-point difference. Other scorers' consensus between two scorings by using ECC has a frequency range between 15 and 40 points. Table 24 also indicates that the percentages of the scores which are the same in two scorings have a range between 2 and 14. This means that the frequencies range between 1 and 6 out of 44 essays. Scorer 8 has no score which is the same for two scorings and the scorers 1, 2, 3, 6, and 9 have only one score which is the same for two scorings. However, Scorer 5 scored 6 essays the same.

For a better understanding of the scorer reliability of the checklist, it is necessary to examine the correlation coefficients between the two scorings done by using ECC. The correlation coefficients computed, by using Pearson Product- Moment Correlation, between the first and the second scorings and they are presented below in Table 25:

Table 25. Correlations across ECC scorings.

Scorer	Correlation coefficient
1	.072
2	.953**
3	.517**
4	.457**
5	.955**
6	.898**
7	.730**
8	.932**
9	.928**
10	.804**

* * Correlation is significant at the 0.01 level

In Table 25, the correlation coefficients range between .072 and .932, which is relatively higher than the correlation coefficients across GIM scorings. Among the ten coefficients, only one of them, which belong to the scorer 1 is not significant. The other correlation coefficients seem significant. This may mean that those scorers gave similar scores to the essays in both scorings. However, 7 of them are above .70 which refer to a high and meaningful correlation coefficient and relatively a high consistency.

III.5. Findings and Discussion of Research Question 5

"Is there any significant difference among the scorer reliability levels of the assessments?"

Although the correlation coefficients between scorings can be significant and the highest range among correlation coefficients were found in scorings done by ESAS, Fischer's z transformation is computed in order to see whether there is a significant difference between different correlation coefficients or not. Therefore, Fischer's z transformations were computed in order to provide a smooth comparison among correlation coefficients across the scorings done by the scorers. The results are given below:

Table 26. The Comparisons using Fisher's z Transformation among Correlation Coefficients across Different Scorings.

		Scorers									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
The Difference between Correlation Coefficients	$r_{12}-r_{34}$	0.056	**p<.05** 2.433	0.099	0.016	**p<.05** 2.992	0.481	0.487	**p<.05** 2.311	1.071	0.498
	$r_{12}-r_{56}$	1.772	0.369	0.282	0.867	1.657	0.702	0.107	0.106	0.109	0.034
	$r_{34}-r_{56}$	1.648	1.572	0.176	0.849	1.282	1.137	0.141	0.205	0.961	0.531

I

n the table, r_{12}refers to the correlation coefficient between the first two scorings done by using GIM; r_{34} refers to the correlation coefficient between the next two scorings done by using ECC; and r_{56}refers to the correlation coefficient between the final two scorings done by using ESAS. The differences at the significant level (p<0,05) are presented in the table. The results indicate that few scorers (2, 5, and 8) did consistent and decisive scorings by using different tools in different time distances. However, this result can be called random because there is not any criteria list used in GIM scorings. As seen in the table, no other consistent and decisive scorings were done by the scorers using different tools in different time distances. This may mean scorers assign different scores to the same essays when they use different assessment tools in different time distances.

III.6. Findings and Discussion of the Research Question 6

"What are the views of the scorers in assessment processes related to the scoring tools?"

Standardized open ended questioning was used for the instrumentation. It includes the same question –the same stimuli- in the same way determined in advance (Patton, 2002). One open-ended question was asked for gathering the views of the scorers on the scorings. The transcripts were analyzed line by line and memos were written (Strauss & Corbin, 1998; Glesne, 1999). Categories or labels were reviewed and recurring themes, core consistencies and meanings were identified by using pattern codes (Miles & Huberman, 1994; Patton, 2002). The themes were found as follows:

a) Criteria Use

b) Spelling

c) Weightings

What is immediately apparent from open-ended transcripts is that the criteria use is very important and useful in essay scoring because the scorers think that they were more precise and reliable in scoring the essays by using the criteria given. One of the scorers states that the scorings done by using general impression marking was like gambling because they needed to assign a total score to each essay. They also state that the criteria use changed attitude to score subjectively in a positive manner. In this respect, scorers agree the idea that scoring by using a checklist or a scale is always more objective and reliable.

Some teachers state that there should be a criterion for spelling. Even if essay writers are in the advanced level, they may make spelling mistakes and the scorers cannot score spelling because it is not one of the criteria in the scale. The spelling criterion was not found appropriate by the judges because the task is an advanced level task. Therefore, they may have not considered that there would be spelling mistakes in the essays which would be written.

Although the scorers accept that the scorings done by using general impression marking were not reliable and consistent, they also criticize the weightings of the criteria in the scale. They state that the criteria should not be equal for each sub-criterion. For example, one of the scorers says it would be better if each weighting was different for each sub-criterion. In this way, it would be more reliable and consistent. Another one says it was not fair to assign the highest score for indentation, pattern variety, or paragraph formatting.

It would be particularly useful to state, considering the transcripts, that criteria use is a reliable measure for assessing essays. However, the criteria should be chosen precisely and correctly, and the weightings of the criteria should be independent from each other.

CONCLUSION

The purpose of this study was to develop a valid and reliable scoring scale which could be used to assess essay writing skill, and to compare the particular scoring scale to general impression marking and checklist tools used in assessing essay writing skill according to scorer/rater reliability. The study gives evidence that all methods, techniques, or tools could include subjectivity and it seems reasonable to notice that mental processes and internal responses of scorers function in different ways in using same scoring criteria for the same essays in different times. In this study, because the purpose was primarily to develop a valid essay criteria checklist and a valid essay assessment scale, and to examine the scoring reliability, different essay writing assessment tools – general impression marking, checklist, and scale – were used for scoring the same essays in different time distances. Those scorings led the way to different but not unexpected results. The results show that general impression marking does make a big difference in scoring essays. The statistical evidence show that the scorings done by using general impression marking is never consistent and reliable. The statistical analyses show that the scorings done by using the checklist (ECC) are more reliable and consistent than the ones done by using general impression marking. The correlation coefficients are higher and they are supported by the scorers themselves with qualitative data. The results also show that the scorings done by using the scale (ESAS) are also consistent and reliable when compared to general impression marking. However, there is a slight difference between the correlation coefficients across ECC scorings and ESAS scorings. Yet, the correlation coefficients across ESAS scorings are slightly higher than the ones across ECC scorings. These slight differences can also be observed by examining the intra-scorer consensus between those scorings. This may mean that different weightings for each sub-criterion may result in more consistent scorings as scorers declared in open-ended questioning because the results of Fischer's z transformation also support the idea that the scores are similar but not the same. This refers, for ECC and ESAS scorings, to different scorings for the same essays in each essay assessment session held in different times. It is very clear that if a scorer assigns the scores 70 and 68 to the same essays in two different sessions and the cut-off score is 70, this means success and failure depend on a source of variation. At this point, the scorers and the time elapsed between scorings may seem as the source of variation.

Bearing in mind that the findings of the assessments in this study should be treated with some caution because of several limitations, further research into the effectiveness and usefulness of the scale would be valuable. It is difficult to infer what processes occurred in scorers' minds while they are scoring essays. In order to obtain verbal descriptions as concrete information, recognize this process, and establish the decision-making processes, think aloud protocols or video recording with follow-up interviews can be employed during the scoring processes of the teachers in order to. In this way, they can be aware of the criteria they take into consideration and develop their own criteria for scorings and this may become a habit in time. This tool can be used at schools for a number of teachers who test the same students in the same way. In fact, think aloud protocols were intended to be employed in this study but the scorers were not eager do it because they found it very difficult and time-consuming. The point to be taken into consideration here is that employing think aloud protocols may need so dedicated and motivated staff to work with.

Another way to assess essays reliably suggested is an automated essay scoring program called e-rater because writing is therefore seen conceptualized as arranging key terms in appropriate sequence. E-rater measures syntactic variety by counting the number of different types of clauses and modal verbs to calculate a ratio of these features per sentence and per essay (Rudner & Gagne, 2001; Herrington & Moran, 2001). Although syntactic variety is not the only feature for assessing essays and because the quality of writing is still based upon criteria outside of the communicative act or content, the tool mentioned can be used for providing syntactic variety itself in essays.

Another way is recommended by Weigle (2002) who states that scorers never agree on writing scores because they bring their own backgrounds, experiences, and values to the assessment. For this reason, an agreement on a set of standards and training, which is for specific group of learners taught, could be helpful for the scorers and it can bring scorers to some agreement.

Because the most important concern in this study is seen as reliable decisions about individuals, the latest developments in measurement and evaluation, Item Response Theory and RASCH measurement also provide some solutions to the reliability limitations. These procedures attempt to reveal the features such as underlying ability of a writer, difficulty or calibration of a test item, scorer leniency or severity, and scorer inconsistency (Bachman, 1990; McNamara, 1996; Tatum, 2000). However, more reliable and consistent decisions about individuals are taken into consideration, Generalizability Theory (G-Theory), instead of correlation, can also be used in order to determine and design reliable scorings, and make accurate decisions for individuals because raters differ in the central tendency and variance, and observations depend on the context in which they occur, and constructs are heterogeneous (Brennan, 2001).

The more pieces of information available, the more reliable will be the conclusions drawn from the data (Cherry & Meyer, 1993). Notwithstanding the limitations, it will be very useful to research and ascertain the usability and effectiveness of ECC and ESAS and a rubric to be developed or in use across different and larger groups of scorers. These studies would help to broaden and improve the conclusions drawn about the reliability and usability of the tools used for scoring essays.

BIBLIOGRAPHY

Aiken, L. R. (1996). *Psychological testing and assessment*. Massachusetts : Allyn and Bacon.

Alderson, J. C., Clapham, C., and Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Anastasi, A. (1997). *Psychological testing*. New Jersey: Prentice Hall Inc.

__________ (1982). *Psychological testing*. New York : Macmillan Publishing.

Atılgan, H., Kan, A., and Doğan, N. (2006). *Eğitimde ölçme ve değerlendirme*. Ankara: Anı Yayıncılık.

Bacha, N. N. (2004). Testing writing in the EFL classroom. *Forum, 40,* 57-64. Retrieved November 11, 2004, http://exchanges.state.gov/forum/vols/vol40/no4/p57.pdf.

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

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

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

Bainer, D. L. and Porter, F. (1992). Teacher concerns with the implementation of holistic scoring. Annual Meeting of the Midwestern Educational Research Association Chicago.

Baldwin, D. (2004). A guide to standardized writing assessment. *Educational Leadership, 62,* 72-75.

Banning, M. (2008). The think aloud approach as an educational tool to develop and assess clinical reasoning in undergraduate students. *Nurse Education Today, 28,* 8-14.

Baykul, Y. (2000). *Eğitim ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Ankara: ÖSYM Yayınları.

Bowen, K. and Cali, K. (2004). *Teaching the features of effective writing.* Retrieved November 21, 2004, http://www.learnnc.org/index.nsf/print View/1216418CB65B73CE85256D7300445C5A?OpenDocument.

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement, 22, 1,* 41-52.

Branch, J. L. (2001). Junior high students and think alouds generating information-seeking process data using concurrent verbal protocols. *Library and Information Science Research, 23,* 107-122.

Breland, H. (1983). *The direct assessment of writing skill: A measurement review* (Technical Report No.83-6). Princeton, NJ: College Entrance Examination Board.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., and Rock, D. A. (1987). *Assessing writing skill.* New York: College Entrance Examination Board.

Brennan, R. L. (2001). *Statistics for social science and public policy: Generalizability theory*. New York: Springer-Verlag Inc.

Brown, H. D. (2001). *Teaching by principles: An interactive approach to language pedagogy*. New York: Addison Wesley Longman, Inc.

Brown, H. D. (1994). *Principles of language learning and teaching*. New Jersey: Prentice Hall Regents.

Brown, J. D. (1995). *The elements of language curriculum.* Massachusetts: Heinle and Heinle Publishers.

Burke, M. J. and Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5(2)*, 159-172.

Carter, R. and Nunan, D. (2001). *Teaching English to speakers of other language*s. Cambridge: Cambridge University Press.

Celce-Murcia, M. (2001). *Teaching English as a second or foreign language*. Massachusetts: Heinle and Heinle.

Chapman, C. (2004). *Authentic writing assessment* (ED328606). Retrieved November 11, 2004, Academic Search Premier database.

Chase, C. I. (1983). Essay test scores and reading difficulty. *Journal of Educational Measurement, 20, 3,* 293-297.

Chase, C. I. (1968). The impact of some obvious variables on essay test scores. *Journal of Educational Measurement, 2, 4,* 315-318.

Chastain, K. (1988). *Developing second language skills: Theory and practice*. New York: Harcourt Brace Jovanovich, Inc.

Cherry, R. and Meyer, P. (1993). Reliability issues in holistic assessment. M. Williamson and B. Huot (Ed.), In *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (p. 109-141). Cresskill, NJ: Hampton.

Clauser, B., Clyman, S., and Swanson, D. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Measurement*, *36*, 29-45.

Coffman, W. (1971). Essay examinations. R. Thorndike (Ed.), In *Educational measurement* (p.271-302). Washington: American Council on Education.

Congdon, P. J. and McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37, 2,* 163-178.

Cooke, L. and Cuddihy, E. (2005). Using eye tracking to address limitations in think aloud protocol. In *International Professional Communication* (195-198). IEEE International Professional Communication Conference New Jersey.

Corder, S. P. (1973). *Introducing applied linguistics*. Aylesbury: Hazell Watson andViney Ltd.

Cotton, D. and Gresty, K. (2006). Reflecting on the think-aloud method for evaluating e-learning. *British Journal of Educational Technology, 37(1),* 45-54.

Croft, K. (1980). *Readings on English as a second language*. New York: Little, Brown and Co.

Cronbach, L., Linn, R., Brennan, R., and Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57,* 373-399.

Cumming, A., Robert K., and Donald E. P. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86,* 67-96.

Darus, S. (2006). Identifying dimensions and attributes of writing proficiency: development of a framework of a computer-based essay marking system for Malaysian ESL learners. *Internet Journal of e-Learning and Teaching, 3(1),* 1-25.

Dempsey, M. S., PytlikZillig, L. M., and Bruning, R. G. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a web-based environment. *Assessing Writing, 14,* 38–61.

De Larios, J. R., Murphy, L. and Manchon, R. (1999). The use of restructuring strategies in EFL writing: A study of Spanish learners of English as a foreign language. *Journal of Second Language Writing, 8(1),* 13-44.

Duran, R. P. (2008). Assessing English-language learners' achievement. *Review of Research in Education, 32,* 292-327.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing, 14,* 88-115.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach in rater variability. *Language Testing, 25,* 155-185.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted rasch model. *Journal of Educational Measurement, 31(2),* 93-112.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33(1),* 56-70.

Ericsson K. and Simon, H. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Erkuş, A. (2006). *Sınıf öğretmenleri için ölçme ve değerlendirme: Kavramlar ve uygulamalar*. Ankara: Ekinoks.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.

Fisher, R., Brooks, G., and Lewis, M. (2002). *Raising standards in literacy*. New York: Routledge.

Funkesson, K. H., Anbacken, E. and Ek, A. (2007). Nurses reasoning process during care planning taking pressure ulcer prevention as an example: A think-aloud study. *International Journal of Nursing Studies, 44*, 1109-1119.

Gamaroff, R. (2000). Rater reliability in language assessment: The bug of all bears. *System, 28(1),* 31-53.

Gerberich, J.R., Greene, H. A., and Jorgensen G. (1963). *Measurement and evaluation in the modern school*. New York: McKay Company.

Glesne, C. (1999). *Becoming qualitative researchers: An introduction*. New York: Longman.

Göransson, K. E., Ehrenberg, A., Ehnfors, M., and Fonteyn, M. (2007). An effort to use qualitative data analysis software for analyzing think aloud data. *International Journal of Medical Informatics, 76,* 270-273.

Gronlund, N. E. (1982). *Constructing achievement tests*. New Jersey: Prentice Hall.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley and Sons Inc.

Gyagenda, I. S. and Engelhard, G. (1998a). Rater, domain, and gender influences on the assessed quality of student writing using weighted and unweighted scoring. Annual Meeting of the American Educational Research Association San Diego.

Gyagenda, I. S. and Engelhard, G. (1998b). Applying the Rasch model to explore rater influences on the assessed quality of students' writing ability. Annual Meeting of the American Educational Research Association San Diego.

Hagen, N. A., Stiles, C., Nekolaichuk, C., Biondo, P., Carlson, L. E., Fisher, K., and Fanesinger, R. (2008). The alberta breakthrough pain assessment tool for cancer patients: A validation study using a delphi process and patient think-aloud interviews. *Journal of Pain and Symptom Management, 35, 2,* 136-152.

Hamp-Lyons, L. (1991). The writer's knowledge and our knowledge of the writer. L. Hamp-Lyons (Ed.), In *Assessing second language writing in academic contexts* (p. 15-36). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1992). Holistic writing assessment for LEP students. Second National Research Symposium on Limited English Proficient Student Studies: Focus on Evaluation and Measurement Washington.

Hartman, H. J., (2001). Pair problem solving and thinking aloud. Retrieved January 10, 2005, http://condor.admin. ccny.cuny.edu/~hhartman /PPS%20&%20TA.html.

Hawkey, R. and Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, *9*, 122-159.

Hays, W. L. (1963). *Statistics*. New York: Holt, Rinehart and Winston Inc.

Heaton, J. B. (1982). *Language testing*. London: Modern English Publishing Lim.

Herman, J., Aschbacher, P., and Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Herrington, A. and Moran, C. (2001). What happens when machines read our students'writing? *College English, 63,* 480-499.

Hieronymous, A., Hoover, H., Cantor, N., and Oberley, K. (1987). *Handbook for focused holistic scoring*. Chicago: Riverside.

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

Hughes, D. and Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement, 21,3,* 277-281.

Hughes, D., Keeling, B., and Tuck, B. F. (1983). Effects of achievement expectations and handwriting quality on scoring essays. *Journal of Educational Measurement, 20(1),* 65-70.

Hughes, D., Keeling, B., and Tuck, B. F. (1980). The influence of context position and scoring method on essay scoring. *Journal of Educational Measurement, 17, 2,* 131-135.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60,* 237-263.

IELTS (2007). *IELTS handbook.* Retrieved January 19, 2008, http://www.ielts.org/_lib/pdf/IELTS_ Handbook_2007 .pdf#.

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

Jaspers, M. W. M., Steen, T., van den Bos, C., and Geenen, M. (2004). The think aloud method: A guide to user interface design. *International Journal of Medical Informatics, 73,* 781-795.

Johnson, R., McDaniel, F., and Willeke, M. (2000). Using portfolios in program evaluation: An investigation of interrater reliability. *The American Journal of Evaluation, 21,* 65-80.

Johnson, R., Penny, J., and Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication, 18(2),* 229-249.

Johnson, R., Penny, J., and Gordon, B. (2000). The relationship between score resolution methods and score reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education, 13,* 121-138.

Johnson, R., Willeke, M., Bergman, T., and Steiner, D. (1997). Family literacy portfolios: Development and implementation. *Window on the World of Family Literacy, 2(2),* 10-17.

Johnstone, C. J., Bottsford-Miller, N. A. and Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kan, A. (2005). Yazılı yoklamaların puanlanmasında puanlama cetveli ve yanıt anahtarı kullanımının (aynı) puanlayıcı güvenirliğine etkisi. *Eğitim Araştırmaları Dergisi, 5, 20,* 166-177.

Kan, A. (2007). Performans değerlendirme sürecine katkıları açısından yeni program anlayışı içerisinde kullanılabilecek bir değerlendirme yaklaşımı: Rubrik puanlama yönergeleri. *Kuram ve Uygulamada Eğitim Bilimleri, 7 (1),* 129-152.

Kane, M., Crooks, T., and Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18(2),* 5-17.

Karmel, L. J. (1970). *Measurement and education in the schools*. Toronto: Macmillan Company.

Keeves, J. P. (1988). *Educational research, methodology, and measurement: An international handbook*. London: Pergamon Press.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioural sciences*. Monterey, CA: Brooks/Cole.

Klein, S. and Hart, F. M. (1968). Chance and systematic factors affecting essay grades. *Journal of Educational Measurement, 5, 3,* 197-206.

Klein, P. D., Piacente-Cimini, S., and Williams, L. A. (2007). The role of writing in learning from analogies. *Learning and Instruction, 17,* 595-611.

Klein, J. and Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. *Assessing Writing, 10,* 134-148.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. New York: Methuen.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26,* 275-304.

Mackenzie, J. (2007). *Essay writing: Teaching the basics from the ground up*. Markham, ON: Pembroke Publishers.

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

Maki, P. L. (2004). *Assessing for learning: Building a sustainable commitment across the institution.* Sterling, VA: Stylus Publishing, LLC.

Marshall, J. C. and Hales, L. W. (1972). *Essential of testing*. Massachusetts: Addison Wesley Publishing.

Marshall, J. C. and Powers, J. M. (1969). Writing neatness, composition errors, and essay grades. *Journal of Educational Measurement, 6, 2,* 97-101.

McNamara, T. (1996). *Measuring second language performance.* London: Longman.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research and Evaluation, 7(25).* Retrieved October 11, 2007, http://PAREonline.net/getvn. asp?v=7andn=25.

Messick, S. (1992). *The interplay of evidence and consequences in the validation of performance assessments.* Princeton, NJ: Educational Testing Service.

Micheels, W. J. and Karnes, M. R. (1968). *Eğitimde başarının ölçülmesi* (Çev. İbrahim Yurt). Ankara: Ajans Türk Matbaası.

Miles, M. B. and Huberman, A. M. (1994). *Qualitative data analysis*. California: Sage Publications.

Miller, M. D., and Crocker, L. (1990). Validation methods for direct writing assessment. *Applied Measurement in Education, 3,* 285–296.

Moskal, B. M. (2000). Scoring rubrics: what, when, and how? *Practical Assessment, Research, And Evaluation, 7(3).* Retrieved October 11, 2007, http://pareonline.net/getvn.asp?v=7&n=3.

Moss, P. A., Gerard, B. J., and Hanniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30,* 109-162.

Moss, P., Schutz, A., and Collins, K. (1997). Developing coherence between assessment and reform: An integrative approach to portfolio evaluation for teacher licensure. Annual meeting of the American Educational Research Association Chicago.

Murphy, K. R. and Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74(4),* 619-624.

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merrill.

Norton, L. S. (1990). Essay-writing: What really counts. *Higher Education, 20(4),* 411-442.

Patton, M. Q. (2002). *Qualitative research and evaluation methods.* California: Sage Publications.

Penny, J, Johnson, R. L., and Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing, 7(2),* 143-164.

Polio, C.G. (1997). Measures of linguistic accuracy in second language writing research. *Language Testing, 19(2),* 109-131.

Prapphal, K. (2008). Issues and trends in language teaching and assessment in Thailand. *Language Testing, 25(1)*, 127-143.

Prime, N. J. and Le Masurier, S. B. (2000). Defining how we think: An investigation of decision making processes in diagnostic radiographers using the think aloud technique. *Radiography, 6,* 169-178.

Raimes, A. (1983). *Techniques in Teaching Writing*. Oxford: Oxford University Press.

Reicks, M., Smith, C., Henry, H., Reimer, K., Atwell, J., and Thomas, R. (2003). Use of the think aloud method to examine fruit and vegetable purchasing behaviors among low-income African American women. *Journal of Nutrition Education and Behavior, 35, 3,* 154-160.

Roberts, V. L. and Deborah I. F. (2006). Methods for inclusion: Employing think aloud methodology in software usability studies with individuals who are deaf. *International Journal of Human-Computer Studies, 64,* 489-501.

Rudner, L. and Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research and Evaluation, 7(26).* Retrieved June 30, 2006, http://PAREonline. net/getvn.asp?v=7 &n=26.

Sasaki, M. and Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing, 16, 4,* 457-478.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25(4),* 465-493.

Schellings, G., Aarnoutse, C., and Leeuwe, J. (2006). Third grader's think-aloud protocols: Types of reading activities in reading an expository text. *Learning and Instruction, 16,* 549-568.

Schmitter-Edgecombe, M. and Bales, J. W. (2005). Understanding text after severe closed-head injury: Assessing inferences and memory operations with a think-aloud procedure. *Brain and Language, 94,* 331-346.

Scott, B. J. and Vitale, M. R., (2003). Teaching the writing process to students with LD. *Intervention in School and Clinic, 38(4),* 220-224.

Simpson, A. (2005). A process approach to writing. *Developing Teachers, 1*. Retrieved January 10, 2005, http://www.developingteachers.com /articlestchtraining/pw1_adam.htm.

Spandel, V. (2005). *Creating writers through 6-trait writing assessment and instruction*. Boston, MA: Pearson Education, Inc.

Stern, H. H. (1983). *Fundamental concepts of language teaching*. Hong Kong: Oxford University Press.

Strauss, A. and Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. California: Sage Publications.

Sulsky, L. M. and Balzer W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73,*497-506.

Tatum, D. S. (2000). Rasch analysis: An introduction to objective measurement. *Laboratory Medicine, 31(5)*, 272-274.

Tekin, H. (1991). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı Kitap ve Yayınevi

________ (2000). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı Kitap ve Yayınevi

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin Company.

Turgut, F. (1997). *Eğitimde ölçme ve değerlendirme metotları*. Ankara: Gül Yayınevi.

Ur, Penny (1996). *A course in language teaching*. Cambridge: Cambridge University Press.

Van den Bergh, H. and Eiting, M. H. (1989). A method of estimating rater reliability. *Journal of Educational Measurement, 26(1),* 29-40.

Van Someren, M. W., Barnard, Y. F., and Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modeling cognitive processes*. London: Academic Press.

Vaughan C. (1991). Holistic assessment: What goes on in the rater's mind? L. Hamp-Lyons (Ed.), In *Assessing Second Language Writing in Academic Contexts* (p. 111-126). Norwood, NJ: Ablex.

Vaughan, D. K. and Farr, D. E. (2004). Performance, education, and experience factors as predictors of writing ability. Retrieved November 28, 2004, http://www.stc.org/ confproceed/1997/PDFs/0127.PDF++%22assessing +writing+ skill%22&hl=t.

Veal, L. R. and Hudson, S. A. (1983). Direct and indirect measures for large-scale evaluation of writing. *Research in the Teaching of English, 17,* 290-296.

Victori, M. (1999). An analysis of writing knowledge in EFL composing: A case study of two effective and two less effective writers. *System, 27,* 537-555.

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16,* 194-209.

Weigle, S. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (1990). *Communicative language testing*. Wiltshire: Prentice Hall.

Welch, C. and Martinovich-Barhite, D. (1997). *Reliability issues and possible solutions*. Annual Meeting of the American Educational Research Association Chicago.

Wexley, K.N. and Youtz, M.A. (1985). Rater beliefs about others: Their effects on rating errors and rater accuracy. *Journal of Occupational Psychology, 58,* 265-275.

White, E. M. (1993). Holistic scoring: Past triumphs, future challenges. M. M. Williamson and B. A. Huot (Ed.), In *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (p. 79-108). Cresskill, NJ: Hampton Press.

Whittington, M. S., Lopez, J., Schley, E. and Fischer, K. (2000). Using think-aloud protocols to compare cognitive levels of students and professors in college classrooms. Annual National Agricultural Education Research Conference San Diego.

Woehr, D. J. and Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67,* 189-205.

Woodford, P. E. (1980). Foreign language testing. *The Modern Language Journal, 64, 1,* 97-102.

Yang, S. C. (2003). Reconceptualizing think aloud methodology: Refining the encoding and categorizing techniques via contextualized perspectives. *Computers in Human Behaviour, 19,* 95-115.

Yancey, K. C. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication, 50(3),* 483-503.

Zamel, V. (1982). Writing: The process of discovering meaning. *TESOL Quarterly, 16,* 195-209.

Zhu, W. (2001). ITRS: A program to compute interrater reliability. *Measurement in Physical Education and Exercise Science, 5(1),* 57-62.

APPENDICES

Appendix 1:

ESSAY CRITERIA CHECKLIST (ECC)

-Make a checkmark if the essay includes the following attributes-

CRITERIA		ATTRIBUTES	CHECKMARK
ORGANIZATION	**A. INTRODUCTION**		
	A.1.1. Introductory Sentences	Effective introductory sentences	
	A.1.2. Thesis Statement	Appropriate thesis statement (thesis and central idea)	
	A.2. BODY PARAGRAPHS		
	A.2.1. Topic Sentence	Appropriate topic sentence (possibly implied) supporting the thesis and the central idea	
	A.2.2. Supporting Sentences	Appropriate sentences supporting the topic (possibly major and minor)	
	A.3. CONCLUSION	Appropriate conclusion related to thesis	
LANGUAGE USE	**B.1. Word Order**	Correct word order	
	B.2. Pattern Variety	Using different patterns	
	B.3. Verb Form	Using verb forms correctly	
	B.4. Tenses	Using tenses appropriately	
	B.5. Articles	Using articles correctly	
	B.6. Pronouns	Using pronouns correctly	
	B.7. Prepositions	Using prepositions correctly (verb + preposition, adjective + preposition)	
VOCABULARY	**C.1. Word Choice**	Selecting the appropriate words	
	C.2. Word Variety	Having a rich vocabulary	
	C.3. Parts of speech	Using the correct parts of speech	
MECHANICS	**D.1. Punctuation**	Using punctuation marks correctly	
	D.2. Capitalization	Using cases (lower/upper) correctly	
	D.3. Paragraphing	Correct paragraph formatting	
	D.4. Indentation	Using margins correctly and consistently	
IDEAS/ CONTENT	**E.1. Title**	Appropriate title	
	E.2. Development	Appropriate development	
	E.3. Unity	Unity	
	E.4. Transitional Signals	Using appropriate transitional signals	

Appendix 2:

ESSAY ASSESSMENT SCALE (ESAS)

CRITERIA		ATTRIBUTES	4	3	2	1	0
ORGANIZATION	B. INTRODUCTION						
	A.1.1. Introductory Sentences	Effective introductory sentences					
	A.1.2. Thesis Statement	Appropriate thesis statement (thesis and central idea)					
	A.2. BODY PARAGRAPHS						
	A.2.1. Topic Sentence	Appropriate topic sentence (possibly implied) supporting the thesis and the central idea					
	A.2.2. Supporting Sentences	Appropriate sentences supporting the topic (possibly major and minor)					
	A.3. CONCLUSION	Appropriate conclusion related to thesis					
LANGUAGE USE	B.1. Word Order	Correct word order					
	B.2. Pattern Variety	Using different patterns					
	B.3. Verb Form	Using verb forms correctly					
	B.4. Tenses	Using tenses appropriately					
	B.5. Articles	Using articles correctly					
	B.6. Pronouns	Using pronouns correctly					
	B.7. Prepositions	Using prepositions correctly (verb + preposition, adjective + preposition)					
VOCABULARY	C.1. Word Choice	Selecting the appropriate words					
	C.2. Word Variety	Having a rich vocabulary					
	C.3. Parts of speech	Using the correct parts of speech					
MECHANICS	D.1. Punctuation	Using punctuation marks correctly					
	D.2. Capitalization	Using cases (lower/upper) correctly					
	D.3. Paragraphing	Correct paragraph formatting					
	D.4. Indentation	Using margins correctly and consistently					
IDEAS/ CONTENT	E.1. Title	Appropriate title					
	E.2. Development	Appropriate development					
	E.3. Unity	Unity					
	E.4. Transitional Signals	Using appropriate transitional signals					
		TOTAL SCORE					