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

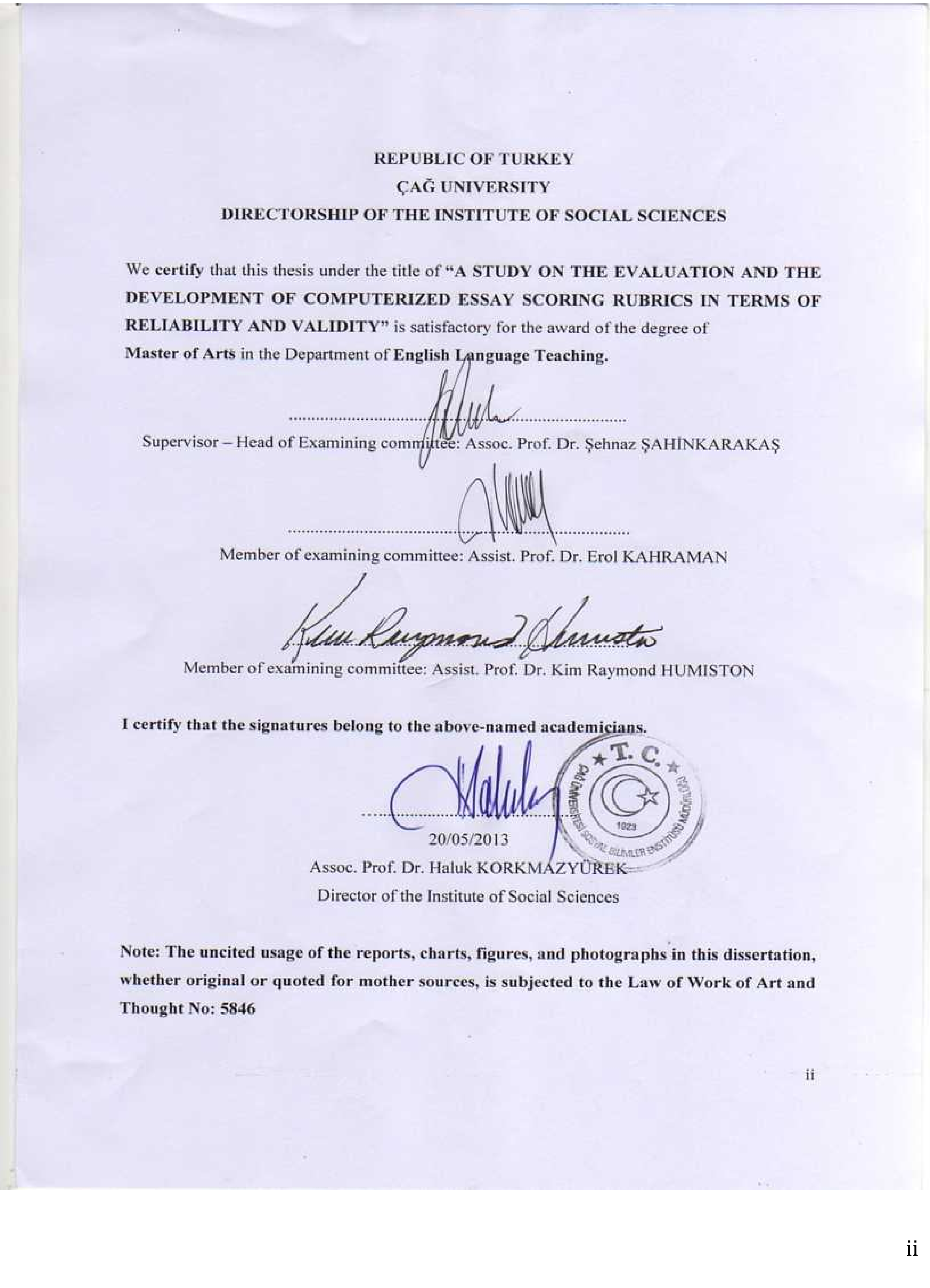
A STUDY ON THE EVALUATION AND THE DEVELOPMENT OF COMPUTERIZED ESSAY SCORING RUBRICS IN TERMS OF RELIABILITY AND VALIDITY

THESIS BY
Hasan SAVAŞ

SUPERVISOR
Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ

MASTER OF ARTS

MERSİN, May 2013



REPUBLIC OF TURKEY
ÇAĞ UNIVERSITY
DIRECTORSHIP OF THE INSTITUTE OF SOCIAL SCIENCES

We **certify** that this thesis under the title of **"A STUDY ON THE EVALUATION AND THE DEVELOPMENT OF COMPUTERIZED ESSAY SCORING RUBRICS IN TERMS OF RELIABILITY AND VALIDITY"** is satisfactory for the award of the degree of **Master of Arts** in the Department of **English Language Teaching.**

....

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

....

Member of examining committee: Assist. Prof. Dr. Erol KAHRAMAN

Kim Raymond Humiston

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

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

Haluk

T. C.
ÇAĞ ÜNİVERSİTESİ SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜ
1923

....

20/05/2013
Assoc. Prof. Dr. Haluk KORKMAZYÜREK
Director of the Institute of Social Sciences

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

ACKNOWLEDGEMENTS

Primarily, I owe almost all my gratitude to my thesis advisor, Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ for her entire support and endless guidance during this study. Without her whole understanding and constructive feedback, it would not be possible to complete this study.

I would like to thank to my thesis committee members, Assist. Prof. Dr. Erol KAHRAMAN, Assist. Prof. Dr. Hülya YUMRU, and Assist. Prof. Dr. Kim Raymond HUMISTON for sharing their knowledge with me all the time.

I am very grateful to my older brother, Muzaffer ÇETİN for being with me at every step of this thesis. For two years, we walked along this road together.

I also would like to send many thanks to my colleagues, Duygu ÖZMEN, Neslihan GÜZEY, and Mihrican YİĞİT for their help and allocating their valuable time with me when I needed. More thanks are to my friends, Ahmet Erdost YASTIBAŞ and Yusuf UYAR for sharing all their support at any time with me.

And I would like to express my love and gratitude to my beloved ones, my mother, my father, and my little sister. They are the lights of my life.

20.05.2013
Hasan SAVAŞ

DEDICATION

To my mother, my father
and my sister.
They are the lights of my life.

ÖZET

YAZMA BECERİSİ PUANLAMA RUBRİKLERİNİN GÜVENİLİRLİK VE GEÇERLİLİK AÇISINDAN DEĞERLENDİRİLMESİ VE GELİŞTİRİLMESİ ÜZERİNE BİR ÇALIŞMA

Hasan SAVAŞ

Yüksek Lisans Tezi, İngiliz Dili Eğitimi Anabilim Dalı
Tez Danışmanı: Doç. Dr. Şehnaz ŞAHİNKARAKAŞ
Mayıs 2013, 91 sayfa

Bu çalışma Zirve Üniversitesi İngilizce Hazırlık Okulu'nda orta ve ileri orta seviyeler için kullanılan yazma dersi puanlama rubriklerinin güvenilirlik ve geçerlilik açısından değerlendirilmelerini ve geliştirilmelerini araştırmaktadır. Rubrikler içerik geçerliliği ve puanlayıcılar arası güvenilirlik açısından değerlendirilmiştir.

Rubriklerin içerik geçerliliklerini ölçmek adına ilk araştırma sorusu olarak beş yazma dersi okutmanının katılımı ile odak grup görüşmesi yapılmıştır. Görüşmenin amacı yazma dersi kazanımları ve puanlama rubriklerinin ne ölçüde birbirleri ile örtüştükleridir. Odak grup, Hazırlık Okulu yazma dersi kazanımları ile rubrik içeriklerinin uyumlu olduğu fakat rubrik maddelerinin tekrar gözden geçirilmesine ihtiyaç duyulduğu sonucuna varmıştır.

İkinci araştırma sorusu için, toplamda 351 C (orta) seviye ve D (ileri orta) seviye öğrenci kompozisyonları rubrik puanlayıcıları arasındaki güvenilirliği ölçmek adına pearson korelasyon katsayısı kullanılarak analiz edilmiştir. Analiz sonuçları göstermiştir ki; öğrenci kompozisyonları toplam skorlarının pearson korelasyon katsayısı sonuçları 0.01 seviyede C seviye için r= .623 ve D seviye için r= .552'dir. Puanlayıcılar arasındaki tutarlılık düşüktür.

Üçüncü araştırma sorusu olarak Zirve Üniversitesi İngilizce Hazırlık Okulu'nda kullanılmak üzere yeni bir kurumsal yazma dersi puanlama rubriği geliştirmek için aynı katılımcılar ile bir odak grup görüşmesi daha yapılmıştır. Odak grup ilk iki araştırma sorusu sonuçlarını göz önünde bulundurarak yeni bir yazma dersi puanlama rubriği geliştirmiştir. Yeni geliştirilen rubrik puanlayıcılar arasındaki tutarlılığı ölçmek için pearson korelasyon katsayısı kullanılarak analiz edilmiştir. Analiz için 59 C (orta) ve D (ileri orta) seviye öğrenci kompozisyonları kullanılmıştır. Analiz sonuçları göstermiştir ki; yeni geliştirilen rubriklerde puanlayıcılar arasındaki güvenilirlik mevcut rubriklere göre daha yüksektir. Öğrenci kompozisyonları toplam skorları için pearson korelasyon katsayısı sonuçları 0.01 seviyede r= .848'dir.

Sonuç olarak, yeni geliştirilen yazma dersi puanlama rubriği Zirve Üniversitesi Hazırlık Okulu'nda kullanılan mevcut rubriklerden daha güvenilir sonuçlar sağlamıştır. Kurumun kazanımları ve ihtiyaçlarına uyumlu kurumsal bir rubrik olmasının beklentileri karşıladığı ve daha tutarlı sonuçlar sağladığı sonucuna varılabilir.

Anahtar Kelimeler: Yazma Becerisi Değerlendirilmesi, Yazma Becerisi Puanlama Rubriği, Güvenilirlik, Geçerlilik

ABSTRACT

A STUDY ON THE EVALUATION AND THE DEVELOPMENT OF COMPUTERIZED ESSAY SCORING RUBRICS IN TERMS OF RELIABILITY AND VALIDITY

Hasan SAVAŞ

Master of Arts, Department of English Language Teaching
Supervisor: Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ
May 2013, 91 pages

This study investigated the validity and the reliability of essay scoring rubrics used for intermediate and upper-intermediate levels at Zirve University English preparatory school. The rubrics were examined in terms of content validity and inter-rater reliability.

In order to determine the content validity of the rubrics, a focus group interview was held with the participation of five writing skill instructors as the first research question. The aim was to what extent the writing class objectives and the descriptors of essay scoring rubrics matched each other. The focus group concluded that the rubrics were compatible with the writing class objectives of the preparatory school, but the descriptors of the rubrics needed to be re-designed.

For the second research question, totally 351 C (intermediate) level and D (upper-intermediate) level students' essays were analyzed by using Pearson r correlation coefficient in order to see the inter-rater reliability between graders of the rubrics. The analysis results showed that the correlation between graders was low as Pearson r results for total scores of the students' essays were r= .623 for C level and r= .552 for D level at the 0.01 level.

As the third research question, one more focus group interview was held with the same participants in order to develop a new institutional essay scoring rubric for Zirve University English preparatory school. The focus group developed a new essay-scoring rubric by taking the results of the first two research questions into consideration. The newly developed rubric was also analyzed by Pearson r correlation coefficient in order to see the inter-rater reliability between graders. 59 C (intermediate) level and D (upper-intermediate) level students' essays were used for that analysis. The analysis results showed that the correlation between graders was higher than the present rubrics as Pearson r results for total scores of the students’ essays were r= .848 at the 0.01 level.

As a result, the newly developed essay-scoring rubric provided more reliable results than the present rubrics used at the preparatory school. It may be concluded that having an institutional rubric, which is compatible with the needs and the objectives of the institution, meets the expectations and provides more consistent grading results.

Keywords: Writing Assessment, Essay Scoring Rubric, Reliability, Validity

ABBREVIATIONS

SPSS: Statistical Package for Social Sciences

LIST OF TABLES

TABLE OF CONTENT

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 1

1. INTRODUCTION

1.1. Background of The Study

Testing and evaluation have always been a significant part of English language learning and teaching environment. As they are at every step of life, extensive evaluations of particular jobs are always needed. In educational environment, testing and evaluation branches work for that. In order to assess students' learning progress, several testing and assessment criteria are applied. Assessment is a key issue to the education process. On this, Mcnamara (2000) makes a common explanation "testing is a universal feature of social life" (p. 3). In that perspective, testing and evaluation world varies in itself such as assessing writing skills, assessing speaking skills, or assessing listening skills. As one of those branches is assessing writing skills and it is the productive side of language learning together with speaking skill, it may not be easy, as thought, to test and evaluate it. Most of the time, language teachers and educators use scoring rubrics to evaluate this productive skill. Holistic and analytic writing scoring rubrics are two of the most commonly used ones. Over these terms, Knoch (2011) highlights the importance as; "in practical terms, teachers and raters need some reference point on which to base their decisions" (p. 91). Mcnamara (2000) also expresses the procedures for effective rubric structure as the following; provided that the rating category labels are clear and explicit, and the rater is trained carefully to interpret them in accordance with the intentions of the tests designers, and concentrates while doing the rating, then the rating process can be made objective (p. 37). In addition to objectivity, Stellmack, Konheim-Kalkstein, Manor, Massey, and Schmitz (2009) put a mark to validity and reliability:

> When used as the basis of evaluating student performance, a rubric is a type of measurement instrument and, as such, it is important that the rubric exhibits reliability (i.e., consistency of scores across repeated measurements) and validity (i.e., the extent to which scores truly reflect the underlying variable of interest). (p. 102)

Two of the basic methods such as reliability and validity may help to evaluate students' writing proficiency levels. As a need to obtain such sustainable criteria, many educational institutions have mainly used holistic and analytic rubrics. Students' writing proficiency levels are assessed through reliable and valid scoring rubrics, which are carefully designed and prepared according to the needs and purposes of institutions and students. Evaluating and grading students' writings have a wide coverage in order to achieve valid and reliable assessment results. This achievement necessitates the need for strong reliable and valid evaluation. As Zirve University English Preparatory School uses analytic rubrics to assess students' writing proficiency, the reliability and validity of those rubrics are going to be taken into consideration in this study. In detail, inter-rater reliability and content validity of the previously mentioned rubrics are going to be evaluated and examined. McNamara (2000) explains analytic rubrics as "analytic rating requires the development of a number of separate rating scales for each aspect assessed. Even where analytic rating is carried out, it is usual to combine the scores for the separate aspects into a single overall score for reporting purposes" (p. 44). In accordance with the usage of analytic rubrics, several researchers also state the reasons why analytic rubrics should be used in writing assessment. They agree on the decision that analytic rubrics provide raters more detailed rating criteria through a variety of dimensions. "The analytic rubric is preferred by so many teachers and language programs, for it includes multiple scales and thus offers a collection of scores instead of only one" (Çetin, 2011, p. 472). Analytic rubrics take the advantage of grading a writing paper from different perspectives such as grading the content, vocabulary, grammar, punctuation and spelling and so on. Analytic rubrics give multiple scores along several dimensions. In analytic rubrics, the scores for each dimension can be summed for the final grade (Stellmack et al., 2009). They also support that idea with the following sentence "at the same time, the analytic rubric provides more detailed feedback for the student and increases consistency between graders" (p. 102). Consistency in grading is also an important part of writing assessment. It helps the grading process maintain reliability in itself. According to Polly, Rahman, Rita, Yun, & Ping (2008), with the aim of supporting students' learning, several ways of using rubrics exist. That can be sustained by establishing the reliability and validity of rubrics. At the same time, that strengthens teachers' trust in their use to support learning and teaching. Intra-rater reliability and Inter-rater reliability provide opportunity for that issue. During this study, inter-rater reliability of the computerized essay scoring rubrics will be examined and correlated as Zirve University English Preparatory School tires to maintain reliability of writing scoring rubrics by making two separate English instructors grade the students' writings. Stellmack et al. (2009) express the necessity and importance of inter-rater reliability as following “when the task of grading students’ papers is divided among multiple graders, and in many undergraduate courses at large universities, high inter-rater agreement is particularly important to achieve uniform grading across sections of the course” (p. 103).

1.2. Statement of The Problem

At Zirve University English Preparatory School, there are four terms in an academic year, and each term is two-month long. In each term, five progress tests, one midterm examination, and one final examination are applied in order to assess and measure students’ language learning progresses. The English learning structure of the program is skill-based. Language learning skills are taught separately which are listening and speaking skills, reading and vocabulary skills, grammar skills, writing skills, and expansive reading skills. All of these different skills are tested and evaluated via previously mentioned progress tests, midterm and final examinations. In terms of testing writing skills of students, three out of five progress tests include writing tests. In addition, in both midterm and final examinations, students take one writing test. Those processes account for five writing tests during a two-month session. As the writing scoring rubrics used for the mentioned tests have not been measured whether they are valid and reliable yet, it is seen as a need to analyze them in terms of their validity and reliability. According to Burke, Ouellette, Miller, Leise, and Utschig (2012), rubrics are used to assess and evaluate knowledge. Without the empirical support for reliability and validity, the value of data collected reduces. For the validity of the writing scoring rubrics used at Zirve University English Preparatory School, no validation analyze has been made so far. On the other hand, two different instructors as the first grader and the second grader grade students’ writings. Although the Preparatory School tries to provide inter-rater reliability by using the same criteria and writing scoring rubrics during grading sessions, it is still a problem to get similar scores between the two graders.

1.3. Aim of The Study

Zirve University English Preparatory School uses analytic rubrics to grade writing. There are four main language levels at the preparatory school. Language levels include: A level (elementary level), B level (pre-intermediate level), C level (intermediate level), D level (upper-intermediate level). In A and B levels, students' writings are graded with analytic rubrics which aim to assess meaningful sentence structures and paragraph writing. Whereas, in C and D levels, students are supposed to be able to write several composition and essay types. This study aims to look at the content validity and the inter-rater reliability of computerized essay scoring rubrics.

1.4. Research Questions

1. What is the inter-rater reliability of computerized essay scoring rubrics used in intermediate and upper-intermediate classes at Zirve University English Preparatory School?
2. What is the validity of computerized essay scoring rubrics used in intermediate and upper-intermediate classes at Zirve University English Preparatory School?
3. How can the validity and the inter-rater reliability of the present rubrics be maintained?

1.5. Operational Definitions

During the study, several writing assessment methods and criteria will be in use in order to conduct the intended research. A clear understanding of what analytic and holistic rubrics are, what reliability and validity terms stand for, or on behalf of these, some brief explanations of analytic rubrics and holistic rubrics are needed. Comer (2011) states that "the use of rubrics can provide opportunities for fostering shared practices" (p. 4). During the assessment process, in order to maintain the grading process organized and consistent, a well-organized grading criteria and carefully designed test items are needed. Çetin (2011) puts an descriptive mark on this issue as "in essay scoring higher inter-rater reliability is obtained when novice raters are required or tend to use to the same rubric whether this be holistic versus holistic or analytic versus analytic" (p. 483). A brief explanation of inter-rater reliability is also defined by Şahinkarakaş (1993), "inter-rater reliability is a way to test how consistent two or more raters are in rating the same writing sample" (p. 2). In addition to the inter-rater reliability, content validity of the writing scorings used at Zirve University English Preparatory School is going to be examined. Focus group interviews are going to be held with the instructors who are in charge of preparing the writing tests and are teaching writing classes at preparatory school. In this perspective, the content validity of the rubrics is going to be examined whether they are in accord with the writing class aims and objectives. The importance of validity comes out with that kind of situation. Jonsson and Svingby (2007) argue on this issue as “basically, validity in this context answers the question: Does the assessment measure what it was intended to measure?” (p. 136). Besides these several terminology, it may be reasonable to express for what purposes focus group interviews will be held. Moskal and Leydens (2000) state the importance of focus group interviews, and note that laying the groundwork for more valid and reliable assessment may necessitate discussions on scoring differences between graders. Appropriate changes to rubrics may be essential.

CHAPTER 2

2. LITERATURE REVIEW

2.1. Testing and Assessment

Testing and evaluation constitute one of the fundamental bases of language teaching. They can be seen as the nature of the area. As testing and evaluation are in every person's life, almost every step that we take in our educational life is evaluated. Testing and evaluation may be counted as a need because every ongoing progress needs to be seen whether it is going well or not. As a basic explanation for that Mcnamara (2000) states, "first, language tests play a powerful role in people's lives, acting as gateways at important transitional moments in education, in employment, and in moving from one country to another" (p. 4). Tests play a fundamental and significant role in deciding what to test and evaluate according to the needs of a particular educational program. It is a need to understand what we test and what criteria we use as they have an impact on both the test takers and the curriculum outcomes. Testing and evaluation should be seen not just as a simple evaluation tool, but also as an indication of the ongoing learning process. According to Fulcher and Davidson (2007), "language testing is all about building better tests, researching how to build better tests and, in so doing, understanding better the things that we test" (p. 19). A well-prepared test can contribute to several aspects in a language program, such as adapting it during an academic term or making rearrangements on it provided that the language program has clear goals and objectives for the future use of the test. In addition, several future planning for a course or a curriculum may be developed, rearrangement of learning and teaching objectives may be decided, or developing a feedback criteria may be constructed. According to Rea-Dickins and Germaine (1993) "evaluation is an intrinsic part of teaching and learning. It can provide a wealth of information to use for the future direction of classroom practice, for the planning of courses, and for the management of learning tasks and students" (p. 3). "It is necessary to evaluate the English language skills of students whose first language is not English in order to place them in appropriate language programs, monitor progress, and guide decisions related to moving from one type of program to another or exiting from language services altogether" (Kopriva, 2008, p. 30). Testing and assessment environment have also sub-titles in themselves. They make up for different purposes, and for several measurements. Basically, in general terms, they are divided into two separate headings, which work for and are used for different purposes in education world. Norm-referenced tests and criterion-referenced tests can be given as the most common examples for types of testing and assessment. Brown (1996) explains this terminology as; norm-referenced tests are used to make comparisons in performance while criterian-references tests are to assess how much of the material or set of skills taught in a course is being learned by students. Placement tests and proficiencey tests may be given as examples for norm-referenced tests. Assessing achievement, diagnosing the progress level of students in a course or program, or deciding whether to promote students to the next level of study may be counted as criterian-referenced tests. Mcnamara (2000) states on this issue as "achievement tests accumulate evidence during, or at the end of, a course of study in order to see whether and where progress has been made in terms of the goals of learning" (p. 6). At preparatory schools of many universities in Turkey, both of previously mentioned types of tests are used and writing tests play an important role in the process. As a first step for a student to start an English preparatory school in Turkey, most of the time, he/she takes a norm-referenced test at the very beginning of the academic year. The student is placed to a level class regarding his/her test results. During the English education process across the academic year, the student takes criterian-referenced tests with the aim of seeing the learning process while promoting to a higher-level class.

2.2. Assessing Writing

Teaching and assessing writing skills (Jonsson & Svingby, 2007) in English language teaching environment are challenging. "Tests to see how a person performs particularly in relation to a threshold of performance have become important social institutions and fulfill a gatekeeping function in that they control entry to many important social roles" (Mcnamara, 2000, p. 3). It is (Heaton, 2003) also an important area in the field to make research. Composition courses expose students to write communicatively while helping them use the grammar and other specific writing skills in their own work (Burke et al., 2012). Regarding the writing skills that are learned through those courses, testing and evaluating branch gains importance. Heaton (2003) states, "in the composition test the students should be presented with a clearly defined problem which motivates them to write" (p. 137). Without providing that students sometimes have difficulty in transferring what they have learned in general English writing courses to academic and disciplinary English writing courses. Upon providing clear objectives, assessment of the students' performance takes place. "There are many different aspects and variables in writing that need to be considered when conducting research" (Knoch, 2011, p. 86). One crucial thing in assessing such performances is reliability of evaluation. The other one is validity of assessment. On this issue, Wiggins (1998) states, "assessment has to be credible and trustworthy, and as such be made with disinterested judgment and grounded on some kind of evidence." "With respect to reliability of evaluation, the more consistent the scores are over different raters and occasions, the more reliable the assessment is thought to be" (Moskal & Leydens, 2000).

Now the question is how to enhance validity and reliability of assessment. To Jonsson and Svingby (2007), "the reliable scoring of performance assessments can be enhanced by the use of rubrics" (p. 130). Andrade (1997) states that "a scoring rubric provides a list of criteria's and helps grade the quality from bad to good for each criterion." According to Sezer (2006), "a scoring rubric enables the grader to evaluate student performance by considering the different aspects of that performance" (p. 5). In other words, a scoring rubric is used to decide on the proficiency level of the learners. To Picket and Dodge (2007), "a scoring rubric has three features such as focusing on the evaluation of the pre-determined goal, using a grading interval to rate performance, and demonstrating certain characteristics of performance according to pre-determined standards." Andrade (1997) points out that: A scoring rubric is a tool that is used frequently by teachers to evaluate performance and by taking teacher expectation into consideration increases student performance. At the same time, a rubric helps teacher decrease the time spent on the evaluation of student work. According to Jonsson and Svingby (2007), "rubrics should be analytic, topic-specific, and complemented with exemplars and/or rater training" (p. 141). During the following sub-titles; rubrics, their importance, usefulness, and basic definitions will be mentioned.

2.3. Rubrics

Rubrics have been a significant part of both evaluation and improvement process of learning. Rubrics may be described as scoring equipment that help to set certain goals and expectations to assign and evaluate either orally or written assignments. For many aspects, their importance cannot be underestimated. Comer (2011) discusses the effectiveness of the rubrics and states that rubrics address several issues such as creating a collaborative atmosphere between teachers with different levels of experience, consistent evaluation of shared learning outcomes, providing timely feedback to students, helping teachers turn into evaluators, and being flexible in assessment approach with the assistance of rubrics. On the effectiveness of rubrics, Allen and Knight (2009) also point the importance of rubrics; "rubrics also provide an effective, efficient, equitable assessment method that can be understood and applied by both student learner and academic assessor" (p. 1). Rubrics serve as understanding the outcomes of student work and the ongoing process of the performance. They may demonstrate both students and teachers what specific points are going on a desired level or vice versa. Weak points may be analyzed directly by the help of rubrics. Lovorn and Rezaei (2011) state on this issue that "rubrics have been developed and used by school systems and teachers for decades in attempts to streamline, clarify, and synthesize evaluative measures" (p. 1). Polly, Rahman, Rita, Yun, and Ping (2008) give a basic definition for rubrics as; "a rubric is a scoring guide which describes the qualities to be assessed in pupils' work or performances." Rubric may exclusively be used to assign grades, evaluate, and provide guidance for students' writing (Andrade, 2005; Stellmack et al., 2009). Rubrics serve for education in many aspects. Most of the time, they help both teachers and learners be steady on their intended study purposes. Rubrics may certainly provide students with intended standards or support teachers define outcomes for a course or a program. In addition, rubrics provide a great deal of benefit both for teachers and students such as providing a better organized, effective feedback or decrease in workload in terms of evaluation. "Rubrics help identify strengths and weaknesses in their work when used to give feedback, and that knowing 'what counts' made grades seem fair" (Andrade & Du, 2005, p. 3). Students and teachers take the advantage of quality work thanks to rubrics. Students are exposed to situations that force them to think on the material they study, at the same time teachers feel that they provide their students with meaningful materials. Retinues of rubrics vary. They may serve for different purposes in terms of long term or short term assignments. In that way, we may count them as teaching materials beside their grading duties. Therefore, it would not be wrong to say that they teach as well as they evaluate. Helping students set their own goals for their studies is another point of rubrics. They also express an additional benefit of improvements in the quality of their work, and less anxiety about assignments. In a study conducted by Andrade and Du (2005), a team of students held several focus group interviews in order to analyze rubric-referenced assessment and gave a shared comment for the topic as; "using rubrics helped them focus their efforts, produce work of higher quality, earn a better grade, and feel less anxious about an assignment" (p. 1).

Another point for rubrics is the quality issue and the importance of rubrics. As can be easily predicted, rubrics occupy a significant position in evaluation of students' work as they bring a quality to assessment. Rubrics are pre-designed assessment tools and serve to many teachers at the same time of evaluation. Grading criteria is ready to use and equality in grading is sustained over rubrics. Thanks to their defined criteria, Peat (2006) discusses the advantage of rubrics for their contributions to the objectivity of writing assessment. Jonsson and Svingby (2007) also add; "nontheless, most educators and researchers seem to accept that the use of rubrics add to the qulity of the assessment" (p. 132). Rubrics also, if used efficiently, may serve as teaching assistants. Providing constant feedback to students or students' seeing their own mistakes may be counted for that. Andrade (2005) puts a point on the issue and tells that teaching with rubrics taught him a lot as an assistant professor. The essence of rubrics depends on how they are created and how they are used. He briefly states:

> It is not just about evaluation anymore; it is about teaching. Teaching with rubrics where it gets good. Issues of validity, reliability, and fairness apply to rubrics, too. We need to worry more about the quality of the rubrics that we use (p. 27).

Apparently, rubrics may meet the expectations of both teachers and students in terms of evaluation, getting affective constant feedback, and assisting students to realize their own weak points in first hand.

Mentioning the clarity and the developmental process of rubrics may be essential. To obtain more valid, user-friendly, and reliable rubrics may necessitate a variety of criteria and study. However, it would be wrong to claim that a best rubric can be designed as several factors such as course objectives, student and teacher expectations, and the essence of assessment criteria. The desired clarity and comprehension of rubrics should be supplied, on the other hand. The rubrics should be clear enough for evaluators to use them efficiently and variables should be appropriate and practical to assess in a given test situation (Bresciani, Oakleaf, Kolkhorst, Nebeker, Barlow, Duncan, & Hickmott, 2009; Mcnamara, 1996; Knoch, 2011).

In terms of developing rubrics and need for them, researchers express countless opinions. Andrade (1997), for example, states the importance of taking professional standards of the program and students' needs into consideration. Knoch (2011) highlights on the development of rubrics and mentions a number of decisions, at the descriptor level, that a rating scale developer has to make by putting them into questions such as: "how many bands should the rating scale have? How will the descriptors differentiate between the levels?" (p. 92). Especially, he underlines the importance of the usefulness of feedback provided to test takers/users. He suggests, "scale descriptors which raters use should have parallel descriptors which can be understood by students and used in subsequent writing performances" (p. 95). He also underlines that after preparing the rating scale, the decisions made by developers should result in the best possible product for the specific context in mind. Weigle (2002) provides a detailed explanation for rubrics, brief summary of which is itemized below, to design and develop. The following items can give a variety of ideas to rubric designers in order to develop valid ones. Before examining the items, it is important to mention that the rubric designers need to decide what type of rubrics they are willing to have. That is, it should be made clear whether an analytic rubric is needed or a holistic one meets the requirements. The researcher states that "it is not enough for a rubric to be clear and explicit: it must also be useable and interpretable" (p. 122), and presents the followings briefly:

- Who is going to use the scoring rubric? For what purposes the rubric will be used. Assessor-oriented or user-oriented scales are two options.

- What aspects of writing are most important and how will they be divided up? The rubric should be developed regarding the needs of the program followed. That is, it needs to be taken into consideration whether it is an academic program or a general English course.

- How many points, or scoring levels, will be used? It is important to keep in mind that the number of descriptors may affect the reliability results and should be decided on what situations (for placement tests or diagnostic tests etc.) they will be used.

> - How will scores be reported? Scores from an analytic rating scale can either be reported separately or combined into a total score. It depends on the test type. For placement purposes, a combined total scoring may be useful. On the other hand, for diagnostic purposes, a separate scoring may be useful for diagnostic information or an accurate picture of test takers' abilities in writing (p. 122).

Holistic and analytic scoring may demonstrate differentiated results and serves for a variety of different purposes. Jonsson and Svingby (2007) provide an overall definition both for holistic and analytic rubrics:

> Two main categories of rubrics may be distinguished: holistic and analytical. In holistic scoring, the rater makes an overall judgment about the quality of performance, while in analytic scoring, the rater assigns a score to each of the dimensions being assessed in the task. Holistic scoring is usually used for large-scale assessment because it is assumed to be easy, cheap and accurate. Analytical scoring is useful in the classroom since the results can help teachers and students identify students' strengths and learning needs (p. 131).

2.3.1. Analytic Rubrics

Analytic scoring, according to Clinard (2011), takes the advantage of several features with specified criteria such as ideas and content, organization, voice, word choice, sentence fluency, and conventions. An analytic rubric (Reineke, 2007) articulates levels of performance for each criterion, so the teacher can assess student performance on each criterion. Weigle (2002) also frames analytic rubrics as "depending on the purpose of the assessment, scripts might be rated on such features as content, organizaiton, cohesion, register, vocabulary, grammar, or mechanics" (p. 114). Lovorn and Rezaei (2011) dramatically summarize the basic structure of analytic rubrics as they compile the procedures that using rubrics in assessment has many benefits; rubrics should be well-designed, topic-specific, analytic, and complemented with exemplars, and they should be institution-specialized while serving for specific purposes and for specific group of students. Sometimes, rubrics on which lots of time has been spent to design may give inconsistent results. In the meanwhile, a poorly constructed essay-scoring rubric may distort an overall judgment. If scores are reasonably consistent across faculty, scoring typically proceeds.

However, if faculties in a department do not feel you cannot use results with confidence, you will want to invest some time in discussions of ratings. If a particular assignment is going to be used to make major decisions about a curriculum or program, ensuring accurate and consistent reporting is important. That may be managed by analytic rubric usage. Knoch (2011) points out a similar opinion on that, “to be able to identify strengths and weaknesses in a learners’ writing and to provide useful feedback to students, an analytic scale is needed” (p. 94). In the process of developing rubrics and constructing their descriptors, views and decisions of teachers may change from one to another. On this dilemma, Knoch (2011) concludes, “raters see content as an important aspect of writing assessment. Without exception, all models specify surface level textual features, like grammar, vocabulary and syntax, as components; they differ, however, in their description of the features” (p. 91). Another advantage of analytic rubrics is that they help both students and teachers see the weak or powerful points of the student work. Weigle (2002) discusses on the phenomenon as “more scales provide useful diagnostic information for placement and/or instruction; more useful for rater training. Raters may read holistically and adjust analytic scores to match holistic impression.” Lastly, Stellmack et al. (2009) suggest an idea on how many descriptors should be included in an analytic scoring rubric and add “as a result, we discovered, that a smaller number of criteria was more practical” (p. 104).

2.3.2. Holistic Rubrics

Clinard (2011) summarizes holistic scorings as the whole picture of writing, generalize one score, efficiency, and reliability. Holistic rubrics are mostly designed to provide an easy and quick way to evaluate students’ work. They generally serve the grader to make an overall evaluation on a task or product that students complete. Most of the time, scoring criteria is based on four or five points scale to decide the success level of student work. Weigle (2002) expresses “in a typical holistic scoring session, each script is read quickly and then judged against a rating scale, or scoring rubric, that outlines the scoring criteria” (p. 112). As they are easy to use and have an overall assessing criterion, holistic rubrics may be seen as being preferred more than analytic rubrics by test makers/users. According to Rudner and Schafer (2002), “when there is an overlap between the criteria set for the evaluation of the different factors, a holistic scoring rubric may be preferable to an analytic scoring rubric” (p. 76). Reineke (2007) also expresses the structure of holistic rubrics as “a holistic rubric does not list separate levels of performance for each criterion. Instead, a holistic rubric assigns a level of performance by assessing performance across multiple criteria as a whole." Sample scoring procedures for a student work that a holistic rubric provides may be listed as the following;

> 4 points – Exceeds criteria: Provides ample supporting detail to support solution/argument. Organizational pattern is logical and conveys completeness. Uses effective language; makes engaging, appropriate word choices for audience and purpose. Consistently follows the rules of standard English.
>
> 3 points – Meets criteria: Provides adequate supporting detail to support solution/argument. Organizational pattern is logical and conveys completeness and wholeness with few lapses. Uses effective language and appropriate word choices for intended audience and purpose. Generally follows the rules of standard English.
>
> 2 points – Progresing criteria: Includes some details, but may include extraneous or loosely related material. Achieves little completeness and wholeness through organization attempted. Limited and predictable vocabulary, perhaps not appropriate for intended audience and purpose. Generally does not follow the rules of standard English.
>
> 1 point – Below criteria: Includes inconsistent or few details which may interfere with the meaning of the text. Little evidence of organization or any sense of wholeness and completeness. Has a limited or inappropriate vocabulary for the intended audience and purpose. Does not follow the rules of standard English. (Assessment Planning with Gloria Rogers, Ph.D.: www.abet.org/assessment.shtml)

Besides having advantages, holistic rubrics may have disadvantages. For one thing, holistic rubrics do not provide a detailed feedback to the users. As holistic rubrics are not capable of providing feedback to each aspect of work that is evaluated, it may be difficult to provide one overall evaluation for students during/after grading.

2.4. Validity

Knowledge, mostly in educational institutions, is often assessed and evaluated using rubrics either designed by instructors or downloaded from public websites. An important concern is that many of these measures lack empirical support for reliability and validity, which reduces the value of data collected with these measures. Most of the time, assessment and evaluation are made with the help of rubric usage. Yet, without the support of reliability and validity for those measurements, the value of data collected may reduce.

Increase in the popularity of rubric usage may be seen as another indicator for more valid and reliable assessment. Language used in the descriptors and exemplars of a rubric maintains a vital point and should help operationalize the attributes and performance criteria. In order to sustain objective and consistent evaluation and assessment results, it is seen as an obligation of a test to be valid and reliable. These terms are divided into categories in themselves depending on the test type. For the validity of the writing scoring rubrics used at Zirve University English preparatory school, content validation methods will be needed. Besides, inter-rater reliability will be measured to see whether there is a consistent correlation between two different graders. According to Payne (2003), "reliability and validity are concerned with the consistency and accuracy of the judgments we make about students and their work."

Validity has different methods such as face validity, content validity, construct validity, and empirical validity. Heaton (2003) provides brief explanations for each term of validity as the following:

> - Face Validity: If a test item looks right to other testers, teachers, maderators, and testees, it can be descibed as having at least face validity. Language tests which have been designed primarily for one country and are adopted by another country may lack face validity. A vocabulary or reading comprehension test containing such words as 'typhoon', 'sampan', 'abacus', and 'chopsticks' will obviously not be valid in East Africa no matter how valid and useful a test it has proved in Hong Kong.

> - Content Validity: This kind pf validity depends on a careful analysis of the language being tested and of the particular course objectives. The test should be so constructed as to contain a representative sample of the course, the relationship between the test items and the course objectives always being apparent.
>
> - Construct Validity: If a test has *construct validity,* it is capable of measuring certain specific characteristics in accordance with a theory of language behaviour and learning. This type of validity assumes the existence of certain learning theories or contructs underlying the acquisition of abilities and skills.
>
> - Empirical Validity: This type of validity is usually referred to as *statistical* or *emperical* validity. This validity is obtained as a result of comparing the results of the test with the results of some criterion measure such as an existing test or the subsequent performance of the testees on a certain task measured by some valid test (p. 159).

Validity is defined as any attempt to show that the content of the test is a representative sample from the domain that is to be tested (Fulcher & Davidson, 2007, p. 6). Comer (2011) argues on this issue as, "a valid method should provide accurate assessment of student learning in connection with desired outcomes, and a reliable assessment should offer fair and consistent evaluation of learning in connection with task requirements and expectations" (p. 1). It is important to ensure that writing assignments reflect clear performance criteria in regards to rubric performance and to assure that writing pre-requisites are carefully integrated throughout the program's curriculum and assessment strategy (Burke et al., 2012). "The development of a scale (or a set of scales) and the descriptors for each scale level are of critical importance for the validity of the assessment" (Weigle, 2002). It seems that there is a need for a study on evaluating the quality of rubrics used at educational institutions. Stellmack et al. (2009) discuss,

> Although rubrics are used frequently in evaluating student writing, little research has focused on assessing the quality of rubrics as measurement instruments (i.e., their reliability and validity). Clearly, it is desirable to establish that a rubric is a valid measure of the variable that one is attempting to evaluate (p. 106).

Well-constructed and carefully designed rubrics with the compatibility to the course objectives should be applied to programs confidently. Everything, on the other hand, may not be totally reliable and valid. On that point, Stellmack et al. (2009) state, “merely using an explicit, carefully developed rubric does not guarantee high reliability” (p. 102).

One of the ways to validate rubrics is the focus group interview method, which is one of the bases of this study. In order to see the content validity of current essay scoring rubrics of Zirve University prep school, focus group interviews will be held by the teachers of the school. For the importance of focus group interviews, many researchers share their contributive ideas to this process. According to Comer (2011), the development process of rubric needs to be done by the teachers employing them, and the specific outcomes and objectives of the learning program should be taken into consideration while applying this. Allen and Knight (2009) see the issue as the collaborative work to develop and validate a rubric grounded in the expectations of academics and professionals. “Thoughtful collaboration with colleagues in the design of any rubric is well-advised because many factors, including the wording and meaning of problems, affect the quality of a measure and its uses” (Burke, et al., 2012, p. 21). Upon the development process of scoring rubrics, Tierney and Simon (2004) take attention to an outstanding matter, “the most challenging aspect of designing rubrics for the classroom is in the language used. Although indicators and exemplars can help operationalize the attributes and performance criteria in rubrics, the choice or wording is still critical” (p. 5). In a study that Bresciani at al. (2009) conducted, development procedures for rubrics were discussed. And a twenty member multi-disciplinary team set out to develop a rubric and they examined existing rubrics available internally. The team considered guidelines used for the review of manuscripts submitted for publication.

2.5. Reliability

In language testing and assessment, reliability is a requirement as it maintains the consistency in testing scores. Munoz (2009) takes it to the subject of this study and points out, “since language assessment is always subject to some degree of variation (i.e. unreliability), ensuring as much reliability as possible becomes a natural concern in specific areas such as writing assessment.” Reliability, in a basic structure, is divided into two terms as intra-rater reliability and inter-rater reliability. The first term is the tendency of a rater to give the same result to a student work at different times and situations and the latter one is the tendency of different raters (Weigle, 2002). According to Jonsson and Svingby (2007), variations in rating might occur either in the consistency of a single rater (intra-rater reliability) or between two or more raters (inter-rater reliability). Several factors might have an influence in the variation of grades. Those two terms are explained briefly in the following titles.

2.5.1 Intra-rater reliability

Intra-rater reliability is the consistency of a single grader scoring the same student work at different times. By maintaining the intra-rater reliability, more consistent testing results may be obtained. Şahinkarakaş (1993) summarized the term as "intra-rater reliability indicates how consistent a single rater is in scoring the same set of essays twice with a specified time interval between the first and second scoring. That is, any particular rater would give the same score on both ratings" (p. 17). Mohan, Miller, Dobson, Harvey, and Sherrill (2000) brings an explanation to the term as "ideally, the interpretation of results should be consistent across individual raters and repeated scorings by the same rater" (p. 473). According to Nutter, Gleason, Jenco, and Christians (1993), intra-rater reliability is "the linear relationship between repeated disease assessments of the same sampling units performed by the same rater or instrument" (p. 809).

2.5.2 Inter-rater reliability

Bresciani et al. (2009) express the need for inter-rater reliability in a study, and state "because the rubric in this study was used by a variety of disciplinary faculty to score student research quality, inter-rater reliability measures are worth investigating." Comer (2011) explains inter-rater reliability as "how well do the various individuals taking part assign the same assessment marks?" (p. 5). Fulcher and Davidson (2007) bring another definition to the term and say "for example, the word 'reliability' can also mean 'trustworthiness'; if an employee shows up every day at work, then attendance is reliable because the employer can trust that the worker will do so" (p. 23). In a case that raters have given high and low scores in a similar pattern over a commonly used event or performance, it means that a high coefficiency is existing (Brown, Glasswell, & Harland, 2004). Many researchers and educational institutions nearly all around the world accept the importance of reliability. Along with the validity, it is seen as an essential part of assessment. It is also defined as the consistency of independent scorers who read and score student writing or other performances. When we try to see the situation from the other side, inter-rater reliability may tell us that as the variability in scores decreases, the trust on rubrics' reliability increases. Lovorn and Rezaei (2011) put a point on that in their study as "as range and variability decreased, the rubric scoring attributes became more reliable and accurate, and thus, the rubric became a more valuable assessment tool" (p. 7).

The rubric, which is in use for the assessment, is also an important issue in reliability. It needs to be rater-friendly and be understandable by raters. Jonsson and Svingby (2007) analyze that briefly "ideally, an assessment should be independent of who does the scoring and the results similar no matter when and where the assessment is carried out, but this is hardly obtainable" (p. 133). Several ways to obtain such a consistency exist. Spearman's correlation or Pearson's correlation co-efficiency r methods are two of them. As pearson's r correlation is the method that is used for this study, it has been taken into consideration. Burke et al. (2012) give information on the usage of the term in their study and state, "pearson's correlation coefficient, r, is calculated as where r is a value between 0 and 1. The closer r is to 1, the higher the correlation, indicating a stronger relationship between the pairings" (p. 17). Stemler (2004) and Brown et al. (2004), on the other hand, provide the commonly assumed calculations as correlation between .55 and .75 are accepted as consistent, and values above .70 are seen as acceptable.

It is important to make rubric graders take pre-training on the development of rubrics as they will participate in the test administration. Pre-training may provide a real benefit to the standardization of the evaluation process. According to Munoz (2009), in order to imporve inter-rater reliability a standardized rubric training is needed with the aim of reducing the effect of individulas and the unsystematic variation of results between different raters as it is usually a problem to maintain consistency between two different graders if the inter-rater reliability and validity studies have not been made yet. Tierney and Simon (2004) argue those issues on behalf of the instructionally usefulness of scoring rubrics because of inconsistencies in the descriptions of performance criteria across their scale levels, and add "for scoring rubrics to fulfill their educational ideal, they must first be designed or modified to reflect greater consistency in their performance criteria descriptors" (p. 1).

Especially, in order to obtain more consistent results during grading sessions, training raters on how to use the scoring rubrics efficiently and make them realize correlation between the course objectives and rubric content are very important. Today, almost every educational institution uses scoring rubrics to maintain a valid, reliable, and objective evaluation of student work. Lovorn and Reazei (2011) put a point on this subject matter "recent studies report that the use of rubrics may not improve the reliability of assessment if raters are not well trained on how to design and employ them efficiently" (p. 1). At Zirve University English preparatory school, the testing office department provides all the instructors a training on how to use writing scoring rubrics at the very beginning of the academic year. Similarly, Stellmack et al. (2009) argue that "the training period in most real-world situations likely would amount to one or two sessions at the beginning of a semester" (p. 106). The researchers also found in their study that participants trained of rubrics submitted scores with increased reliability. Assumedly, with careful training, inter-rater reliability could be higher. Therefore, it would not be wrong to come to such a conclusion that it is necessary to train graders to maximize inter-rater reliability agreement and to yield more consistent results (Burke et al., 2012; Zimmaro, 2004). Çetin (2011) basically puts a final point on this issue and concludes, "the data suggests that as far as essay scoring is concerned, if more than one rater are assigned to the assessment of the same essays, then raters' random use of rubrics will most possibly result in rather low inter-rater reliability" (p. 481).

CHAPTER 3

3. METHODOLOGY

3.1. Introduction

At Zirve University English preparatory school, analytic rubrics are used to assess students' writings. The rubrics are prepared by a group of English instructors (writing coordinator, preparatory program coordinator, testing office members, and several writing class instructors) working at the preparatory school. Yet, no study or statistical measurement has been made on the validity and reliability of the rubrics. This study basically aims to investigate the content validity and inter-rater reliability of writing scoring rubrics used at the institution.

3.2. Participants

At the preparatory school, about 120 English instructors are working, and about 1400 university students take English classes during an academic year. Students take a placement test at the very beginning of the academic year. Regarding the previously decided level scores in the test, they get into the level classes such as elementary level (A level), pre-intermediate level (B level), intermediate level (C level), and upper-intermediate level (D level). Each level is two-month long, and that accounts for 8 months-long English learning and teaching environment. A skill-based system is used to teach English at the school, and these skills are taught and assessed separately. Writing classes, grammar classes, listening & speaking classes, reading & vocabulary classes, English laboratory classes, and expansive reading (readers) classes constitute the preparatory school system. Students attend 27 hours of English classes in a week, and number of class hours in a day varies to the essence of the level.

For the first research question, 211 C level students' and 140 D level students' essays were used to assess the inter-rater reliability measurements. The students' ages and genders were not taken into consideration; however, they may be accepted as adult learners. The results of the measurements will be discussed in the following chapter. In order to see the content validity of the rubrics, which is the second research question, a focus group interview was held. The writing coordinator, a testing office member, two writing instructors, and the researcher himself attended the gathering. All the focus group interview participants accepted to take part in the study. The interview was video-recorded in order to be used during the study. Consistency between the writing class objectives and analytic rubric descriptors were discussed. At the end, the problematic descriptors in the rubrics were analyzed. For the third research question, one more focus group interview was held, and it was also video-recorded. The discussions were based on the inter-rater reliability results of the present rubrics and the previous focus group interview's final decisions. Following these periods, developmental discussions were made on the scoring rubrics. As the consequence of the focus group interview, a new analytic essay-scoring rubric was developed. The newly developed scoring rubric was analyzed in terms of inter-rater reliability. A brief information about the participants of the focus group interviews are demonstrated in Table 1:

Table 1. Participants of the focus group interviews

	The Institution worked	Year(s) of experience	The skill(s) taught
Participant 1	Zirve University	4 years	Writing Instructor, Testing Office Member
Participant 2	Zirve University	6 years	Writing Coordinator
Participant 3	Zirve University	4 years	Writing Instructor
Participant 4	Zirve University	3 years	Writing Instructor
Participant 5	Zirve University	6 years	Writing Instructor, The Researcher

3.3. Instruments

To collect data, first and second grades of 351 essays (211 C level and 140 D level) and the present rubrics that are used to grade them, two focus group interviews, the newly developed rubric according to the interviews, and 59 essays used to measure the reliability of the newly developed rubric were used. The present rubrics, focus group interviews, and the newly developed rubric are explained in detail under this chapter.

3.3.1. The Present Rubrics

The present scoring rubrics are analytic in general, but holistic in part. C and D Level rubrics differ from each other in terms of the number of the paragraphs and indents in the essays (three paragraphs and indents in C Level, and five paragraphs and indents in D Level). The aim of the rubrics is to analyze students' essays by grading 14 separate descriptors. The names of the descriptors are 'holistic, title, format, indent, introduction paragraph, body, conclusion paragraph, content, grammar, vocabulary, transitions/connectors, punctuation, spelling, and coherence.' The contents of each descriptor are explained in brief below (For the details of the rubric, see Appendix 1):

1. Holistic: The graders evaluate the student essay holistically by reading it generally, but not in detail.
2. Title: The graders read the essay and evaluate the title of the essay depending on what is written.
3. Format: The graders count the number of the paragraphs in the essay whether there are three paragraphs in C level and five paragraphs in D level.
4. Indent: The graders count the number of the indents (three indents in C level and five indents in D level) in the essay.
5. Introduction paragraph: The graders evaluate the introduction of the essay by the features of introductions studied in writing classes.
6. Body: The body part of the essays is evaluated according to the expression of ideas and to what extent the ideas support the thesis statement.
7. Conclusion paragraph: The conclusion paragraph is evaluated in terms of whether the thesis statement and topic sentences are restated.
8. Content: The graders check whether the ideas are explained well or not.
9. Grammar: The use of grammar in the essay is checked.
10. Vocabulary: Whether or not students use suitable vocabulary for the given topic is checked.
11. Transitions/Connectors: Each essay type has different transitions to mark the supporting ideas. Therefore, the use of transitions and connectors like however, and, or so is checked in terms of whether they are used meaningfully and grammatically correct.
12. Spelling: The correct use of spelling is checked.
13. Punctuation: The correct use of punctuation is checked.
14. Coherence: Whether the essay explains one idea in a coherent way is checked.

3.3.2. Focus Group Interviews

Two focus group interviews were held during the study. The first one was a content analysis for the second research question. As it was video recorded, the recordings were used as the instruments. The second focus group interview was held for the third research question. It was video recorded, and the recordings were used as the instruments.

3.3.3. The Newly Developed Rubric

The newly developed rubric was used as an instrument as the last step of the study. Needed reliability measurements were applied to it by using 59 student essays, and the results will be provided in the following chapter. (For the details of the rubric, see Appendix 2).

3.4. Data Collection Procedures

On 26th February 2013, needed permissions about the study and the focus group interviews were taken from the ethical committee of Zirve University. First focus group interview was held on 1st March 2013 with the attendance of previously stated English instructors. It was at the faculty building during the work-hours after school. The writing class objectives and the rubric descriptors were given to the group members ten days before the gatherings. The aim in that was to provide some time to the participants to get prepared for the focus group interview. The inter-rater reliability results were analyzed between the dates 11th February and 24th February 2013. Following the analysis of the inter-reliability results and focus group interview decisions, on 9th April 2013 one more focus group interview was held in an aim to develop a more valid, reliable, and user-friendly rubric for Zirve University English preparatory school.

3.5. Data Analysis

In order to analyze the inter-rater reliability results of the present rubrics between two graders, C level and D Level students' essays were used. Totally 351 students' essays (211 C level and 140 D level) were analyzed by using Pearson product-moment correlation coefficient (Pearson r) in SPSS 2012. Pearson r correlation coefficient was applied both to the first research question and to the third research question. For the first one, the present rubrics that were prepared by the prep school were used. The aim was also to see how the correlation existed between two graders for those rubrics. For the third research question, the newly developed rubric designed by the focus group members were used to evaluate 59 C and D level students' essays, and the inter-rater reliability of the rubric was measured by Pearson r correlation coefficient. Two focus group interviews were held: one for the second research question and one for the third research question. Each of the focus group interviews was video recorded. And those video recordings were used as the instruments of this study. The first focus group interview was held with the aim of finding an answer to what extent the writing class objectives and rubrics' content were matching each other. The meeting was a 50-minute discussion and mainly about finding the problematic parts of the rubric descriptors. The discussions were in Turkish. The researcher and the participants made comments on the content of the analytic rubrics and of the descriptors one by one. As the rubrics had 14 different descriptors, participants tried to find out how they could work and how the scoring value of them could be appropriate to the assessment criteria. For the inter-rater reliability measurements, each descriptor of the rubrics was analyzed one by one to see whether there was a/any consistency of the descriptors between graders. In addition to descriptors, final scores of the essay gradings were analyzed to see how consistent the rubrics were as a whole. The last focus group interview was held in order to develop a more valid and reliable rubric. The meeting was about 90 minutes long. The discussions were again in Turkish. The researcher and the participants tried to find out new solutions to develop new rubrics. And some parts of the video recordings were translated into English as transcripts in order to be used for the findings of the study. The results and the detailed data analysis are going to be discussed in the following chapter.

3.6. Conclusion

In this chapter, the participants of the study, the instruments that were used, data collection procedures and tools, and how the data analyses were made have been demonstrated basically. The assessment system of Zirve University English preparatory school has also been introduced.

CHAPTER 4

4. FINDINGS AND DISCUSSION

4.1. Introduction

This study investigates to find out what the content validity of the writing scoring rubrics is, to see whether there is inter-rater reliability between graders, and how the content validity and inter-rater reliability of those rubrics may be maintained. In this chapter, findings and discussions for each research question are going to be explained in details.

4.2. Findings and Discussions of Research Question I

In the first research question, I investigated the inter-rater reliability of computerized essay scoring rubrics used in intermediate and upper-intermediate classes at Zirve University English Preparatory School.

Table 2. Pearson correlations for the C and D level rubrics for 'Holistic descriptor'

	Level C	Holistic1	Holistic2
Holistic1	Pearson Correlation	1	.433*
	Sig. (2-tailed)		.000
	N	211	211
Holistic2	Pearson Correlation	.433*	1
	Sig. (2-tailed)	.000	
	N	211	211
	Level D	Holistic1	Holistic2
Holistic1	Pearson Correlation	1	.386*
	Sig. (2-tailed)		.000
	N	140	140
Holistic2	Pearson Correlation	.386*	1
	Sig. (2-tailed)	.000	
	N	140	140

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

The *Holistic scoring* in the rubrics occupies a 10 out of 100 points in C level and a 12 out of 100 points in D level (Table 2). They try to evaluate students' general writing skills as a whole. There are eight variables in these descriptors and all have holistic scorings like zero point, one point, six or ten points and so on (see Appendix 1 and 2). Pearson r results for these descriptors are r= .433 for C level and r= .386 for D level, which show that there is a nearly no correlation between two graders for the holistic scoring in the essays at the 0.01 level. Graders may not be sure what the descriptors try to measure because the statements in these sections may not be so clear enough for holistic scoring. And the variation of descriptors may have a negative effect on consistency between graders. Having a holistic evaluation in an analytic rubric can also be open to discussion.

Table 3. Pearson correlations for the C and D level rubrics 'Title descriptor'

	Level C	Title1	Title2
Title1	Pearson Correlation	1	.706*
	Sig. (2-tailed)		.000
	N	211	211
Title2	Pearson Correlation	.706*	1
	Sig. (2-tailed)	.000	
	N	211	211
	Level D	Title1	Title2
Title1	Pearson Correlation	1	.729*
	Sig. (2-tailed)		.000
	N	140	140
Title2	Pearson Correlation	.729*	1
	Sig. (2-tailed)	.000	
	N	140	140

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

The scoring for *the Title* (Table 3) in the rubrics occupies a two out of 100 points in C level and one point for D level. It is used just to measure whether there is a title (two points or one point) or not (no point) in the essay. Pearson r results for these descriptors are r= .706 for C level and r= .729 for D level, which show that there may be a slight correlation between two graders for the title scoring in the essays at the 0.01 level. Although there are two variables in the descriptor, it is still not totally correlated. This may be because graders grade the title not for its existence, but for its quality. And this may cause inconsistency in grading.

Table 4. Pearson correlations for the C and D level rubrics 'Format descriptor'

	Level C	Format1	Format2
Format1	Pearson Correlation	1	.864*
	Sig. (2-tailed)		.000
	N	211	211
Format2	Pearson Correlation	.864*	1
	Sig. (2-tailed)	.000	
	N	211	211
	Level D	Format1	Format2
Format1	Pearson Correlation	1	.462*
	Sig. (2-tailed)		.000
	N	140	140
Format2	Pearson Correlation	.462*	1
	Sig. (2-tailed)	.000	
	N	140	140

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

Pearson r results for *the Format* descriptors in Table 4 are r= .864 for C level and r= .462 for D level. It indicates a positive correlation between two graders at the 0.01 level for C level. This descriptor has a two-points value in C level and a three-points value in D level in the total scoring and measures the student writing whether there is enough number of paragraphs in the essay or not. Students get total point when they have the right number of paragraphs in their papers in this section. This section may be counted as consistent for C level when looked at the correlation results; however, there is a poor correlation for D level essays. That may be because some D level essays may be confusing for the grader in terms of number of paragraphs in the essays. Some D level essays may seem to the graders as not having correct number of paragraphs although they have.

Table 5. Pearson correlations for the C level rubric 'Indent descriptor'

	Level C	Indent1	Indent2
Indent1	Pearson Correlation	1	.462*
	Sig. (2-tailed)		.000
	N	211	211
Indent2	Pearson Correlation	.462*	1
	Sig. (2-tailed)	.000	
	N	211	211

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

For Table 5, Pearson r result is as r= .462 for C level. The descriptor provides almost no correlation between two graders at the 0.01 level. The graders are needed to give scores to the student's writing if there is enough number of indents in the essay or not. It has two points out of 100 in the rubric. As it can be deduced from the result that grading *Indent* in a rubric may be challenging for a grader. Some students have indents in their papers, but handwriting may be deceptive. In D level rubric, *Indent* is not graded for the reason that we, as the instructors, assume D level students to be capable of having the right number of indents in their essays. Also, it is not seen as a need to grade the indent in upper-intermediate level.

Table 6. Pearson correlations for the C and D level rubrics 'Introduction descriptor'

	Level C	Introduction1	Introduction2
Introduction1	Pearson Correlation	1	.396*
	Sig. (2-tailed)		.000
	N	211	211
Introduction2	Pearson Correlation	.396*	1
	Sig. (2-tailed)	.000	
	N	211	211
	Level D	Introduction1	Introduction2
Introduction1	Pearson Correlation	1	.377*
	Sig. (2-tailed)		.000
	N	140	140
Introduction2	Pearson Correlation	.377*	1
	Sig. (2-tailed)	.000	
	N	140	140

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

Correlation results in Table 6 (*Introduction*) show that graders may have difficulty in understanding the aim. Pearson r result as r= .396 for C level and r= .377 for D level at the 0.01 level indicate that some graders may not exactly be sure what to grade. This section of the rubric has an eight-point value and tries to make it clear whether student has covered the right steps such as a hook sentence or a correct thesis statement in an introduction paragraph. There is nearly no correlation between graders. In here, graders may have thought that introduction paragraph should be graded for its grammatical or mechanical features although the aim of the rubric is to grade content. The descriptors may be re-arranged according to the needs of the writing class objectives of the preparatory school.

Table 7. Pearson correlations for the C and D level rubrics 'Body descriptor'

	Level C	Body1	Body2
Body1	Pearson Correlation	1	.366*
	Sig. (2-tailed)		.000
	N	211	211
Body2	Pearson Correlation	.366*	1
	Sig. (2-tailed)	.000	
	N	211	211
	Level D	Body1	Body2
Body1	Pearson Correlation	1	.211*
	Sig. (2-tailed)		.013
	N	140	140
Body2	Pearson Correlation	.211*	1
	Sig. (2-tailed)	.013	
	N	140	140

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

In scoring *Body* paragraph(s) in Table 7, the descriptors occupy 10 out of 100 points. These sections of the rubrics try to measure whether there is enough number of topic sentences and based on that whether there are coherent examples or details. For instance, student gets four points in this descriptor if 'There is a body paragraph with some missing details/examples'. However, it is 10 points provided that 'There is a well developed body paragraph'. Pearson r results for these descriptors are r= .366 for C level and r= .211 for D level at the 0.01 level, which indicate nearly no correlation between two graders while scoring body paragraph(s) in the essays. This may be because the rubric training at the very beginning of the academic year is not so effective and informative. Graders who do not teach writing classes at the preparatory school may not be sure what the descriptors in here try to measure as the statements in this section do not provide enough guidance. Redesigning the descriptors that include the intended information in them may be reasonable.

Table 8. Pearson correlations for the C and D level rubrics 'Conclusion descriptor'

	Level C	Conclusion1	Conclusion2
Conclusion1	Pearson Correlation	1	.260*
	Sig. (2-tailed)		.000
	N	211	211
Conclusion2	Pearson Correlation	.260*	1
	Sig. (2-tailed)	.000	
	N	211	211
	Level D	Conclusion1	Conclusion2
Conclusion1	Pearson Correlation	1	.189*
	Sig. (2-tailed)		.025
	N	140	140
Conclusion2	Pearson Correlation	.189*	1
	Sig. (2-tailed)	.025	
	N	140	140

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

The scoring *Conclusion* paragraphs (Table 8) in the rubrics have an eight out of 100 points value. These descriptors are used to look for conclusion, restatement of the thesis statement, and a closure sentence. Students get three points in this descriptor if, for example, 'they have a bad conclusion.' However, it is eight points for 'a well-developed ending/conclusion.' Pearson r results for these descriptors are r= .260 for C level and r= .189 for D level at the 0.01 level. It is so clear that there is almost no correlation between two graders for conclusion paragraphs in the essays. The reason of low correlation may be because the rubrics do not serve the graders with the correct aim of the conclusion paragraph. As it is in the introduction paragraphs, graders may perceive conclusion paragraphs as sections to evaluate grammar and mechanics although the aim is to look at the content.

Table 9. Pearson correlations for the C and D level rubrics 'Content descriptor'

	Level C	Content1	Content2
Content1	Pearson Correlation	1	.343*
	Sig. (2-tailed)		.000
	N	211	211
Content2	Pearson Correlation	343*	1
	Sig. (2-tailed)	.000	
	N	211	211
	Level D	Content1	Content2
Content1	Pearson Correlation	1	.472*
	Sig. (2-tailed)		.000
	N	140	140
Content2	Pearson Correlation	.472*	1
	Sig. (2-tailed)	.000	
	N	140	140

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

Scoring *the Content* of the essays in the rubrics (Table 9) occupies a 10 out of 100 points. Student papers are graded in terms of meaning, coherence, enough number of topic sentences, or details/examples in the whole essay. However, descriptors in this section may not provide the needed criteria as the Pearson r results for these descriptors are r= .343 for C level and r= .472 for D level at the 0.01 level. The correlation between two graders for these sections is not high. The reason for that may be because descriptors are so general; for example, student gets four points in this descriptor if 'There is a composition, but not enough quality' or seven points for a 'satisfactory quality'. Graders may again not be certain what the content of an essay tries to measure as the descriptors look like holistic scoring. Clearer explanations to the descriptors may be needed.

Table 10. Pearson correlations for the C and D level rubrics 'Vocabulary descriptor'

	Level C	Vocabulary1	Vocabulary2
Vocabulary1	Pearson Correlation	1	.326*
	Sig. (2-tailed)		.000
	N	211	211
Vocabulary2	Pearson Correlation	.326*	1
	Sig. (2-tailed)	.000	
	N	211	211
	Level D	Vocabulary1	Vocabulary2
Vocabulary1	Pearson Correlation	1	.200*
	Sig. (2-tailed)		.018
	N	140	140
Vocabulary2	Pearson Correlation	.200*	1
	Sig. (2-tailed)	.018	
	N	140	140

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

Scoring *Vocabulary* in the rubrics (Table 10) occupies a 10 out of 100 points. It tries to measure students' level-specified vocabulary knowledge. It is two points if students 'have some mistakes on the basic vocabulary they know, and there are errors in meaning.' On the other hand, students get 10 points 'with a creative use of vocabulary for the level.' Pearson r results for these descriptors are r= .326 for C level and r= .200 for D level, which show that there is a nearly no correlation between two graders for scoring vocabulary in the essays at the 0.01 level. The term for level-specified vocabulary may not mean so much for graders as they teach different skills at preparatory school.

Table 11. Pearson correlations for the C and D level rubrics 'Grammar descriptor'

	Level C	Grammar1	Grammar2
Grammar1	Pearson Correlation	1	.297*
	Sig. (2-tailed)		.000
	N	211	211
Grammar2	Pearson Correlation	.297*	1
	Sig. (2-tailed)	.000	
	N	211	211
	Level D	Grammar1	Grammar2
Grammar1	Pearson Correlation	1	.448*
	Sig. (2-tailed)		.000
	N	140	140
Grammar2	Pearson Correlation	.448*	1
	Sig. (2-tailed)	.000	
	N	140	140

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

Table 11 (*Grammar*) occupies a 12 out of 100 points. Level-specified grammar structures are evaluated in these sections. Grading criteria is diverse such as 'no correct sentences / no points', 'OK but some problems / six points', or 'very good, nearly no mistakes / twelve points'. Graders have five different descriptor options to grade, and that may be the cause of inconsistency in scoring between graders. Pearson r results for these descriptors are r= .297 for C level and r= .448 for D level, which indicate nearly no correlation between two graders while scoring grammar in the essay at the 0.01 level. The diversity of descriptors may be the cause for inconsistency between graders. For this reason, need for rubric training again may have a crucial importance as Reynolds-Keefer (2010) states in his/her study that an important consideration in the use of rubrics is that training teachers and administrators and modeling the use of rubrics may increase the likelihood those rubrics will be used in the future.

Table 12. Pearson correlations for the C level rubric 'Punctuation descriptor'

	Level C	Punctuation1	Punctuation2
Punctuation1	Pearson Correlation	1	.037*
	Sig. (2-tailed)		.591
	N	211	211
Punctuation2	Pearson Correlation	.037*	1
	Sig. (2-tailed)	.591	
	N	211	211
	Level D	Punctuation1	Punctuation2
Punctuation1	Pearson Correlation	1	.186*
	Sig. (2-tailed)		.027
	N	140	140
Punctuation2	Pearson Correlation	.186*	1
	Sig. (2-tailed)	.027	
	N	140	140

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

Pearson r results for *Punctuation* (Table 12) are r= .037 for C level and r= .186 for D level at the 0.01 level. The scoring value for punctuation in the rubrics is three points. There is no doubt that it would be ridiculous to talk about consistency between graders with these correlation results. It may not be easy for graders to be consistent while grading due to diversity of items in these descriptors.

Table 13. Pearson correlations for the C and D level rubrics 'Spelling descriptor'

	Level C	Spelling1	Spelling2
Spelling1	Pearson Correlation	1	.295*
	Sig. (2-tailed)		.000
	N	211	211
Spelling2	Pearson Correlation	.295*	1
	Sig. (2-tailed)	.000	
	N	211	211
	Level D	Spelling1	Spelling2
Spelling1	Pearson Correlation	1	.155*
	Sig. (2-tailed)		.068
	N	140	140
Spelling2	Pearson Correlation	.155*	1
	Sig. (2-tailed)	.068	
	N	140	140

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

Pearson r results for *Spelling* (Table 13) are r= .037 for C level and r= .155 for D level at the 0.01 level. The scoring value for spelling in the rubrics is three points. As it is the same with punctuation, there is no consistency between graders. There are four descriptors in this section, and graders seem to have difficulty in deciding what criteria suits best to the student paper. It seems that it could be a better idea to reduce the number of descriptors in the rubric and to re-design the items by considering the learning outcomes of the target language level.

Table 14. Pearson correlations for the C and D level rubrics 'Transitions descriptor'

	Level C	Transitions1	Transitions2
Transitions1	Pearson Correlation	1	.174*
	Sig. (2-tailed)		.011
	N	211	211
Transitions2	Pearson Correlation	.174*	1
	Sig. (2-tailed)	.011	
	N	211	211
	Level D	Transitions1	Transitions2
Transitions1	Pearson Correlation	1	.317*
	Sig. (2-tailed)		.000
	N	140	140
Transitions2	Pearson Correlation	.317*	1
	Sig. (2-tailed)	.000	
	N	140	140

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

Scoring *Transitions and Connectors* (in Table 14) occupy a 10 out of 100 points. Students are expected to use certain transitions and connectors such as however, on the other hand, besides and so on. To give some examples, students get two points for 'little use of transitions, many problems with the meaning of the connectors /transitions' or eight points for 'appropriate use of transitions, few problems with the meaning'. Pearson r results for these descriptors are r= .174 for C level and r= .317 for D level, which provide nearly no correlation between two graders at the 0.01 level. The reason for the inconsistency is most probably the high number of descriptors, which causes variation in scoring between graders.

Table 15. Pearson correlations for the C and D level rubrics 'Coherence descriptor'

	Level C	Coherence1	Coherence2
Coherence1	Pearson Correlation	1	.332*
	Sig. (2-tailed)		.000
	N	211	211
Coherence2	Pearson Correlation	.332*	1
	Sig. (2-tailed)	.000	
	N	211	211
	Level D	Coherence1	Coherence2
Coherence1	Pearson Correlation	1	.386*
	Sig. (2-tailed)		.000
	N	140	140
Coherence2	Pearson Correlation	.386*	1
	Sig. (2-tailed)	.000	
	N	140	140

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

Scoring *Coherence* in the rubrics (in Table 15) occupies a 10 out of 100 points. It tries to measure students' coherent ideas in their essays. No point is given to the paper in a case that the essay has 'no coherence/paragraph has more than one idea'. It is ten points provided that the paper has 'good coherence and one idea.' Pearson r results for these descriptors are r= .332 for C level and r= .386 for D level. Correlation between graders for both descriptors is low at the 0.01 level.

Table 16. Pearson correlations for the C and D level rubrics 'Total scores'

	Level C	Total1	Total2
Total1	Pearson Correlation	1	.623*
	Sig. (2-tailed)		.000
	N	211	211
Total2	Pearson Correlation	.623*	1
	Sig. (2-tailed)	.000	
	N	211	211
	Level D	Total1	Total2
Total1	Pearson Correlation	1	.552*
	Sig. (2-tailed)		.000
	N	140	140
Total2	Pearson Correlation	.552*	1
	Sig. (2-tailed)	.000	
	N	140	140

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

Pearson r results for *Total Scores* of the students' essays (in Table 16) are r= .623 for C level and r= .552 for D level. Correlation between graders is low at the 0.01 level. The total results show that a re-arrangement and designing on the scoring rubrics may be made in order to obtain a more reliable one.

It may be concluded that inter-rater reliability between graders of C and D level writing scoring rubrics are not high at Zirve University English preparatory school. An analytic rubric with less descriptors and with clearer scoring shame may be a need for the institution.

4.3. Findings and Discussion of Research Question II

For the second research question, we investigated the validity of computerized essay scoring rubrics used in intermediate and upper-intermediate classes at Zirve University English Preparatory School.

As the scoring criteria and the type of writing are mostly similar for D level and C level writing classes, only the C level rubric and the descriptors inside the present rubric were discussed during the focus group interview. The rubrics and their descriptors are presented in Appendix 1 and 2. Focus group interview results are also provided in this chapter.

Before having discussions on the rubrics, the writing class objectives for C and D levels were provided to the participants by the researcher. The objectives are presented below:

Zirve University Prep School Writing Objectives and Goals

Description of the course: Writing is one of the two skills that students produce the language they learn and communicate in their foreign/second language. Writing classes give the opportunity to the students to express themselves in English to their teachers and peers in written form while writing classes of course is a part of their learning process throughout the prep year.

Main Goal: By the end of the year, students will be able to write academic essays at upper-intermediate level clearly and accurately at an acceptable speed. They will also have a critical awareness of their writing in terms of content, coherence, and linguistic accuracy.

Level C

Main goal: Students will be able to write three-paragraph compositions on a given topic and develop an idea in an organized way.

Objectives: By the end of level C, a student;

- can write an introduction, body and conclusion paragraph.
- can express the main idea in a well-written thesis statement.
- can give enough background information about the topic.
- can write a hook to get readers' attention.
- can write about a person

- can provide examples
- can narrate a personal experience
- can supply reasons
- can support the main idea with parallel points
- can follow process writing steps (brainstorming, outlining, writing first draft, editing, writing final draft)

Level D

Main goal: Students will be able to write five-paragraph essays with appropriate rhetorical organization.

Objectives: By the end of Level D, a student;

- can convey information
- can express reasons
- can develop an argument
- can make comparisons between two event, objects and so on.
- can state a problem and offer solutions
- can present arguments for or against a position

As a start to the focus group interview, the researcher made an entrance speech to the participants in order to remind them of the aim of the meeting. Several issues were stated. During the meeting, the essay scoring rubrics (the present rubrics), which were in use at that time at the preparatory school, were going to be the main topic. The aim was to discuss on the rubrics to what extent the content of the rubric descriptor matched the writing class objectives. As the rubrics were analytic and had 14 different descriptors, the descriptors were going to be discussed one by one. There were two separate rubrics for C and D levels; however, they were almost the same. They only differed in terms of the number of paragraph formats. C level necessitates a three-paragraph essay while D level does a five-paragraph essay. For this reason, one of the rubrics (C level) was taken into consideration during the focus group interview.

The results of the focus group interview and discussions for each descriptor are provided as the following:

The first descriptor of the rubric 'Holistic'

It was argued whether *Holistic Scoring* is a need or not in an analytic rubric. It was also stated that there were eight descriptors in this section and scoring criteria were very close to each other. That may cause the grading to lose its consistency between graders. The following arguments were made on the holistic scoring:

> Participant 5: In an analytic rubric, we have a *Holistic* descriptor, and personally, I do not see that it is necessary. In short sentences and limited details in the descriptor, there is something missing for a holistic view. It does not measure the rubric as a whole, I guess.
>
> Participant 3: I think so. Statements such as 'can be improved or in a good condition' are not totally holistic in my opinion. They are also very close to each other meaningfully.
>
> Participant 1: In this descriptor, it is not clear whether it measures students' grammar knowledge or the content of the essay. Every descriptor is open to comments.
>
> Participant 5: Furthermore, I think *Holistic rubric* is extra in an analytic rubric like this one.

The present rubrics are already analytic rubrics, and may not be a need to have holistic scoring in them. As the focus group stated, holistic scoring seems not to have clear statements. The items are not totally clear for deciding an exact score, and may disturb the consistency between graders. As a consequence, the last decision was that *Holistic Scoring* should either be taken out of the rubric or be taken into consideration for a change in the next focus group interview.

The second descriptor of the rubric 'Title'

On this descriptor, no much discussion was made. It was stated that students needed to know how to add a title to the essays in C and D levels. Therefore, just scoring the existence of the title is good. There is no need to make changes on this descriptor.

Participant 5: This item can stay in the rubric as it is.

Participant 3: Yes, students know that they have to add a title to their writings.

It is open to question whether it is necessary to score *Title* in the rubric or not; however, students are taught to include a title in their essays. That is why; scoring *title* may stay as it is in the rubric or may be inserted into another descriptor.

The third and fourth descriptors of the rubric 'Format' 'and 'Indent'

Nearly the same decisions with the *Title* section were taken for the *Format and Indent* sections. It was suggested to talk on them during the next focus group interview in which developmental discussions were going be made on the rubrics.

Participant 2: It is not necessarily important to spend too much time on these descriptors, in my opinion.

Participant 3: I agree with you. These descriptors are easy to grade. The grader just needs to count the right number of paragraphs and indents.

As a result, deciding on the *Format and Indent* was left for the next focus group interview. It was decided that they totally matched the writing class objectives. They are also part of writing classes in prep school and are emphasized during classes.

The fifth descriptor of the rubric ‘Introduction Paragraph’

The variables of the *Introduction Paragraph* were analyzed one by one. Several discussions were made as the following:

Participant 4: The scoring of the descriptor is not bad. The thesis statement, background information and the hook may be added to the descriptor in order to make it clearer for the graders.

Participant 3: I agree with this idea.

Participant 5: More details, as you stated, may be added to the descriptor. That would help the graders have a better understanding on the rubric.

Participant 2: Graders need to know that the thesis statement means or what is a hook? Without that, adding more details would not be effective.

As a result, it was decided that adding clearer details, which are more comprehensible for graders, to *th*e *Introduction* section could be reasonable. Besides, graders need to be informed of that scoring introduction paragraph in the rubric is about content of the essay not grammar or structure. That may be taken into consideration.

The sixth descriptor of the rubric ‘Body’

In *Body Paragraph*, topic sentences and the supporting details are graded in terms of their development and coherence. In order to find the problematic parts, several decisions were taken. The participants provided the following comments:

Participant 2: To what extent do the descriptors such as ‘many problems’ or ‘missing details’ provide us the needed scoring criteria?

Participant 5: It is not clear whether those are for grading mechanics or content.

Participant 2: Some more emphasis can be put over the meaning. That is, more meaningful and clear expressions can be added to them.

Participant 4: While doing that, we need to consider that descriptors are not long for graders to read. Graders may be bored while reading them if we add more details to the descriptors.

Participant 5: During the next meeting, we may take the aim of the body paragraph into consideration while developing new descriptors.

As the final decision, topic sentences and supporting details about them could be added to the new rubric; however, participants should consider the descriptors would be reader-friendly in order not to bore the graders.

The seventh descriptor of the rubric 'Conclusion Paragraph'

For *the Conclusion Paragraph*, the descriptors are similar to introduction paragraph. Therefore, it was decided that the same procedures would be applied to it during the next meeting. The participants made following comments:

Participant 3: As in the introduction paragraph, the importance of thesis statement should be expressed. It should be clear that restatement of the thesis statement is included in the descriptors.

Participant 4: I agree. What do 'bad conclusion' and 'OK (not good, not bad)' stand for? Those descriptors should be made clearer.

As scoring the conclusion of an essay is similar to scoring introduction paragraph, the importance of scoring should be on the content of the essay.

The eighth descriptor of the rubric 'Content'

The aim of this section (*Content*) is to grade students' writing in terms of meaningful explanation of the ideas and the main idea. In order to make this descriptor more valid, following decisions were made:

Participant 2: This descriptor tries to assess the length of the paragraphs as well as the content.

Participant 4: One of the descriptors is; 'There is a composition, but not enough quality' or 'satisfactory'. It does not sound clear to me. What is satisfactory in here? It does not make sense, in my opinion.

Participant 5: Yes, you are right. Some importance may be given to the content of the paragraphs.

Participant 3: Yes. The descriptors in here should remind the graders of the supporting details and topic sentences.

It is clear and understood from the comments that more meaningful statements needed to be added to the descriptor. Expressions such as 'not enough quality' or 'satisfactory' did not provide enough explanation to grade the content of the essay. While developing the new rubric at the next focus group interview, those final decisions could be taken into consideration.

The ninth descriptor of the rubric 'Vocabulary'

For *the Vocabulary* section, the following conversations were made:

Participant 4: For this section, the descriptors are not problematic, I guess. The problem is that our system is skill-based system, and writing graders don not totally know what kinds of vocabulary are taught in reading and vocabulary classes. Therefore, while grading vocabulary, it is not clear what 'basic vocabulary only' or 'Appropriate word choice but no surprises. Uses level vocabulary in appropriate places' in the descriptor stand for. Reading and vocabulary coordinator may provide the graders a list of vocabulary that is taught in the classes. In that way, graders may be aware of the target vocabulary and grade students' papers by taking them into consideration.

Participant 3: Then, I am sure nobody cares about that list. It is also time consuming for graders.

Participant 2: Actually, most of the graders have classes with the students, so they are aware of what level they are grading. The issue is also the correct use of part of speech in vocabulary.

Participant 5: On the other hand, the next descriptor is transitions and connectors. They may also be seen as vocabulary. For the new rubric, combining these two descriptors may be possible. The items, at the same time, may be decreased.

For *the Vocabulary* section, it was decided that the items were not so clear enough for the graders and they needed to be rearranged according to the C and D levels' objectives. The number of items may also be reduced with an aim to provide more consistency between graders.

The tenth descriptor of the rubric 'Grammar'

During the conversations on this descriptor, it was argued that whether there should be more emphasis on grammar usage or not. On the other hand, it was realized that C and D levels-specified grammar knowledge was not indicated. Graders need to be informed of the grammatical structures of the mentioned levels in the rubric. The following conversations were made on that issue:

Participant 2: C and D level grammatical structures should be included in the rubric.

Participant 4: Well, you mean that this descriptor can provide the grammatical structures such as relative clauses and passive voice in the descriptor, don' you?

Participant 2: Yes, but this may be confusing for the graders, at the same time. In these levels, students need to mostly be able to use complex and compound sentences rather than simple sentences. Grading should be based on that.

Participant 5: Descriptors are also not so clear in here. What does 'many problems' mean? What does it measure? I agree with your idea.

Participant 2: I think, all the descriptors in this section should be re-designed in the next meeting.

Final decision was to redesign *the Grammar* section in the next meeting by considering the C and D level grammar requirements. In that way, a more valid rubric may be obtained.

The eleventh and twelfth descriptors of the rubric 'Punctuation' and 'Spelling'

There is no difference in the rubric in terms of grading *Punctuation and Spelling*. Descriptors share the same statements and the marking shame. During the conversations on this issue, it was stated that both punctuation and spelling are significant points in writing. Participants provided several ideas on that as the following:

Participant 4: Some students have problems in punctuation and spelling. They need to be graded for these writing areas separately.

Participant 2: They are indispensible parts of writing, I think.

Participant 4: Some student papers are great in content, but have noting correct in punctuation or spelling. Descriptors in our rubric are clear enough to grade them particularly.

Participant 3: Graders may be informed of the correct usage of punctuation and spelling.

Punctuation and spelling should be emphasized in the rubric. Students are required to write essays that are structurally in good condition. On the other hand, for the new rubric taking them into one title can be an option.

The thirteenth descriptor of the rubric 'Transitions/Connectors'

The shared idea on this section (*Transitions/Connectors*) was that the descriptors were perfectly matching with the writing class objectives of the preparatory school. Similar discussions with the previous descriptors were made on this section; for example, what can be understood from an expression like 'some use of transitions, some problems'? or what does 'appropriate use of transitions, few problems with the meaning' mean in this descriptor'? Participants gave several opinions on those as the following:

Participant 2: In most of the writing papers of students, I do not see much use of transitions. That is, keeping a separate section for transitions and connectors may not be necessary.

Participant 5: I told you before that I still insist on combining these two sections (vocabulary and transition/connectors) into one descriptor.

Participant 2: This issue is open to dispute. Students, while they are not obliged to, need to use transitions/connectors in order to be able to write comprehensible sentences.

Participant 5: Think that a student has used a few transitions/connectors in his/her essay, but the meaning is clear. On the other hand, another student has used lots of transitions/connectors in his/her essay, but most of them were wrongly used. So to what extent will we grade this section?

Participant 2: Yes, you are right.

Participant 5: These descriptors need to be reconsidered during the next meeting.

Because a variety of transitions and connectors are taught and promoted to students in writing classes, this section satisfies the needs of the writing class objectives. However, combining this section with the vocabulary descriptors may be argued.

The fourteenth descriptor of the rubric 'Coherence'

In this section (*Coherence*), students' expressing one idea and being connected to the topic sentence are graded. Participants stated that there was no need to make changes in this descriptor. Ten-point scale is OK for coherence. Some ideas were provided for *the Coherence* section as the following:

Participant 5: Everything seems clear in this section. I just think that second descriptor is similar to the first one.

Participant 2: the second one may be stated more clearly.

Participant 5: That, for instance, may be as 'Some problems with the coherence while expressing one idea'.

Participant 2: Right. 'No coherence' means that there is more than one idea in the essay.

Participant 4: The second item tries to measure that student has one idea in his/her essay, yet the supporting details are not coherent.

Participant 5: It is not stated in the second item, in my opinion. It should be more understandable for the grader.

As it is stated in the discussions, the second item seems problematic. During the next focus group interview, adding more comprehensible statements to that descriptor was taken as the final decision. In that way, graders could easily understand what the rubric measures.

The overall reflection of the first focus group interview can be that the rubrics and their descriptors mostly match the writing class objectives of the preparatory school; however, the number of descriptors in several items are too much and may cause the graders spend extra time while grading. That can also affect the inter-rater reliability between two graders negatively. Besides, the content of several descriptors are not clear enough, and that can cause the grader not to be able to decide on the right choice.

4.4. Findings and Discussions of Research Question III

The third research question was to find out the ways to maintain the validity and inter-rater reliability of the currently used rubrics.

By taking the final decisions of the first focus group interview into consideration, another one was held with the same participants. The Pearson r results of the present rubric were shown to the participants, and the problematic descriptors were discussed together. Besides, the participants were reminded of examining the final decisions of the former focus group interview before the meeting. In that way, no time was spent to talk over them again. The researcher started the meeting by demonstrating a variety of analytic rubrics that are used by several other educational institutions. Those sample rubrics had already been sent to the participants via e-mail before the meeting. The meeting was about 50 minutes long and several decisions were taken in order to design a new rubric. The first decision was that the new rubric should have maximum four to five descriptors as the present one had 14 descriptors and had poor inter-rater reliability results. On this issue, Mertler (2001) argues "if a rubric contains four levels of proficiency or understanding on a continuum, quantitative labels would typically range from '1' to '4'." One of the well-known analytic rubrics in English language assessment environment is the ESL Composition Profile prepared by Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey (1981). Becker (2011) states on the rubric as the following:

> This rubric, which provided the first conceptualization of scoring separate components for writing, consisted of five major analytic dimensions (i.e., development, organization, vocabulary, language use, and mechanics) designed to measure the writing of ESL students at North American universities (p. 115).

With the purpose of designing a new rubric, participants shared several ideas as the following:

> Participant 5: The Person r results are clear. We need to decrease the number of descriptors without losing the content validation in order to sustain the inter-rater reliability between the graders. We also need to take the writing objectives into account.

> Participant 1: The sample analytic rubrics that we have seen today support that idea, I guess.

According to Weigle (2002), the number of descriptors needs to be designed carefully by examining a variety of rubrics, and taking the needs of the curriculum into account is also important. Upon discussing on the present rubric's descriptors, it was decided to combine them in five headings as *Essay Format, Content, Organization, Use of Language, and Vocabulary*, and appoint them appropriately weighted scoring points. In that way, participants tried to design a user-friendly rubric without sacrificing the content validity. The issue of having an institutional rubric and its meeting the expectations of writing class objectives was also taken into consideration. On this, Johnson and Vosmik (2007) report "we note that faculty using the rubric should assign weighted point values to each section when grading, based on their own course expectations and goals" (p. 7).

Under *the Essay Format*, it was decided to evaluate the student work in terms of the number of paragraphs and indents. *Essay Format* was also used to evaluate student whether he/she had a title or not in his/her essay. Five out of 100 points were given to this descriptor.

Content was intended to evaluate student's work according to its appropriateness to the topic and coherence of the essay. 30 out of 100 points were given to the content section.

Organization was designed for introduction, body, and conclusion paragraphs. This section aimed to evaluate student's work whether the essay had a thesis statement, topic sentences and supporting details about them. In conclusion paragraph, it was intended to measure students' essays in terms of a closure sentence and re-statement of the thesis statement. 25 out of 100 points were given to this descriptor.

Together with *the Content* section, *Organization* occupied 55% of the total scoring in the rubric. In that way, 45% of the scoring was for grammatical and mechanical evaluation. The newly developed rubric and its descriptors are provided in Appendix 3.

In *the Use of Language* section, evaluation of grammar, punctuation, and spelling were brought together. The newly developed rubric tried to assess student's essay whether those terms we equal to the intended level of the student. 25 out of 100 points were given to this descriptor.

In *Vocabulary* section, transitions, connectors, and vocabulary were agreed on to be included. Student's essay was evaluated for level-specified vocabulary and necessary transitions and connectors for a certain essay in this section. 15 out of 100 points were specified for this descriptor.

The newly developed rubric was shared with the focus group interview participants via e-mail. All the participants agreed on the descriptors and the scoring criteria. In order to see the inter-rater reliability of the rubric, 59 randomly selected C (intermediate) and D (upper-intermediate) level students' essays were used. Eight instructors as the first graders and the second graders graded the essays by using the newly developed essay-scoring rubric. In order to assess the inter-rater reliability results, Pearson r was applied to the scoring results. Correlation results for each descriptor and the total scores are demonstrated in the following tables:

Table 17. Pearson correlations for the newly developed rubric 'Essay Format descriptor'

		Essayformat1	Essayformat2
Essayformat1	Pearson Correlation	1	.551*
	Sig. (2-tailed)		.000
	N	59	59
Essayformat2	Pearson Correlation	.551*	1
	Sig. (2-tailed)	.000	
	N	59	59

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

The Essay Format scoring (Table 17) in the newly developed rubric occupies a five out of 100 points. It aims to evaluate students' writings in terms of the number of paragraphs and indents, and whether there is a title or not. The Pearson r correlation result is r= .551 at the 0.01 level, which indicates a poor correlation between the first and second graders of the essays. In the present rubrics, title, indent, and format sections are used separately to measure what *Essay Format* in the newly developed rubric tries to measure. The Pearson r results of *Title, Format, and Indent* in the present rubrics are r= .706, r= .864, and r= .462 in C level, and r= .729 and r= .462 in D level in order; however, *Indent* is not included in level D rubric. In content validity, the items in the present and newly developed rubrics are valid as they are relevant to the writing objectives, but the discussion results of the second focus group interview report that *Essay Format* in the new rubric is not time-consuming, and is user-friendly as the graders are not supposed to re-read the essays two and three times to evaluate when compared to the present rubrics.

Scoring *the Content* (Table 18) in the newly developed rubric occupies a 30 out of 100 points. This descriptor intends to evaluate student's writing according to its appropriateness to the topic and coherence of the essay. The Pearson r result is r= .703 at the 0.01 level, which indicates a more positive correlation between the first and second graders of the essays when compared to the previous correlation results of the present rubrics. Hamp-Lyons (2003) states "writing tests cannot be 100%, and are rarely more than 80%, reliable." Thus, having a correlation result such as r= .703 can be seen as reliable.

Table 18. Pearson correlations for the newly developed rubric 'Content descriptor'

		Content1	Content2
Content1	Pearson Correlation	1	.703*
	Sig. (2-tailed)		.000
	N	59	59
Content2	Pearson Correlation	.703*	1
	Sig. (2-tailed)	.000	
	N	59	59

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

Table 19. Pearson correlations for the newly developed rubric 'Organization descriptor'

		Organization1	Organization2
Organization1	Pearson Correlation	1	.721*
	Sig. (2-tailed)		.000
	N	59	59
Organization2	Pearson Correlation	.721*	1
	Sig. (2-tailed)	.000	
	N	59	59

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

Pearson r result for *the Organization* section (Table 19) is r= .721 at the 0.01 level, which indicates a more positive correlation between the first and second graders of the essays when compared to the correlation results of the present rubrics. This descriptor occupies a 25 out of 100 points in the newly developed rubric. In the present rubrics, this section is scored as *Introduction Paragraph, Body Paragraph, Conclusion Paragraph* separately, and the content of these paragraphs. Pearson r results for those were not reliable, at all. Having just one descriptor for organization has nearly two times more reliable results with the newly developed rubric. That may indicate that having clear and shorter explanations in a rubric provide the grader decide on the score more accurately. The test developers may need to carefully decide on the descriptors. Pitoniak, Young, Martiniello, King, Buteux, & Ginsburgh (2009) state:

> Item writers and reviewers should work to ensure that all test items maintain specificity in their match to content guidelines. As part of the process of creating and reviewing test material to ensure that it is appropriate and accessible to examinees, it is important that item developers, state content review staff, and state review committees analyze each item critically to ensure that it only measures the intended construct (p. 12).

Table 20. Pearson correlations for the newly developed rubric 'Use of Language descriptor'

		Useoflanguage1	Useoflanguage2
Useoflanguage1	Pearson Correlation	1	.840*
	Sig. (2-tailed)		.000
	N	59	59
Useoflanguage2	Pearson Correlation	.840*	1
	Sig. (2-tailed)	.000	
	N	59	59

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

Pearson r result for *the Use of Language* (Table 20) is r= .840 at the 0.01 level. That may mean a positive correlation between the first and the second graders. This descriptor occupies a 25 out of 100 points in the newly developed rubric. In the present rubrics, this section is divided into three descriptors as *Grammar, Punctuation, and Spelling*; however, in the newly developed rubric it is scored for one descriptor in the name of *Use of Language*. It can be referred that combining mechanical components of the essay and grading them in just one descriptor maintain inter-rater reliability.

Pearson r result for *the Vocabulary* section (Table 21) is r= .717 at the 0.01 level, which indicates a more positive correlation between the first and second graders of the essays when compared to the same section's correlation results of the present rubrics. This descriptor occupies a 15 out of 100 points in the newly developed rubric. In the newly developed rubric, vocabulary is graded as a whole under one descriptor instead of scoring transitions, connectors, and vocabulary separately as it is in the present rubrics. In that way, graders have chance to see the vocabulary usage as a whole together with transitions, connectors, and vocabulary. This way also helps the rubric provide more reliable results.

Table 21. Pearson correlations for the newly developed rubric 'Vocabulary descriptor'

		Vocabulary1	Vocabulary2
Vocabulary1	Pearson Correlation	1	.717*
	Sig. (2-tailed)		.000
	N	59	59
Vocabulary2	Pearson Correlation	.717*	1
	Sig. (2-tailed)	.000	
	N	59	59

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

Table 22. Pearson correlations for the newly developed rubric 'Total Scores'

		Total1	Total2
Total1	Pearson Correlation	1	.848*
	Sig. (2-tailed)		.000
	N	59	59
Total2	Pearson Correlation	.848*	1
	Sig. (2-tailed)	.000	
	N	59	59

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

Pearson r result for *Total Scores* in Table 22 is r= .848 at the 0.01 level. That may mean a positive correlation between the first and the second graders for the total scores of the newly developed rubric. Pearson's correlation results of the present rubrics, for which there is nearly no correlation between graders, are provided in findings and discussion of the first research question. It can be inferred from the correlation results of both the present rubrics and the newly developed rubric that having a decreased number of descriptors in a rubric and providing clearer items at the same time help the rubric provide more reliable results.

4.5. Conclusion

In this chapter, the content validity of the present rubrics was discussed, and the inter-rater reliability results of those rubrics were analyzed. Based on the results, a new essay scoring rubric for Zirve University English preparatory school was developed with the help of focus group interviews. The purpose was to have a valid and a reliable institutional essay-scoring rubric for intermediate and upper-intermediate level English classes in the preparatory school. Having an institutional rubric that serves for the needs of the educational program of the institution is important. Hamp-Lyons (2003) argues on this subject matter as "we have increasingly come to realize that local development and implementation, when done well, is a powerful force for positive educational change."

CHAPTER 5

5. CONCLUSION

5.1. Overview of the Study

The purpose of this study was to look at the content validity and inter-rater reliability of writing scoring rubrics used at Zirve University English preparatory school. The study also aimed to develop a valid and reliable writing-scoring rubric in order to use at the intermediate and upper-intermediate levels at the preparatory school. For the validation of the rubrics, focus group interviews were held. Correlation analysis between two graders of the rubric was calculated with the purpose of finding and maintaining the inter-rater reliability. The study addressed three research questions.

In order to find the results for the first research question, 211 C level students' and 140 D level students' essays were used to assess the inter-rater reliability measurements. Pearson r was conducted to the essay scorings. The overall implication of the analysis may be that there is a poor correlation between the graders for the present essay scoring rubrics.

For the second research question, a focus group interview was held and it was discussed to what extent the present rubrics, used at Zirve University English preparatory school, match the objectives of the writing classes. Eventually, the focus group participants shared the idea that the present essay scoring rubrics mostly meet the needs of writing class objectives of the preparatory school; however, several descriptors in the rubrics may not be entirely satisfying.

For the last research question, a new essay scoring rubric, which is analytic again, was developed without losing the content validity. After that in order to see the inter-rater reliability of the newly developed rubric, 59 C (intermediate) level and D (upper-intermediate) level students' essays were used. The results were more reliable when compared to results of the present rubrics.

5.2. Discussion of Findings

According to Jonsson and Svingby (2007), "since performance assessments are more or less open ended per definition, it is not always possible to restrict the assessment format to achieve high levels of reliability without sacrificing the validity" (p. 141). While developing the new rubric for the preparatory school, the maximum effort was made by the focus group interview participants to avoid going out of the frame of writing class objectives. All the focus group interview participants agreed on each descriptor of the new rubric and stated the new rubric meets the expectations of the institution. However, inter-rater reliability results of the rubric were not still totally reliable. Nevertheless, the results were higher than the present rubrics' results. On this issue, Johnson and Vosmik (2007) state that a rubric was developed in a study they conducted. They tried to maintain the content validity and inter-rater reliability by having several meetings with their commitee. However, the inter-rater reliability results were not at the level that they expected. Therefore, they concluded:

> Our follow-up discussions and reflection on content validity led to another revision of the rubric to include 18 topical categories and more clearly operationalized research skills. We also more clearly differentiated among research skills and communication skills. We scored two additional papers, but our inter-rater reliability was still under 70% (p. 8).

An important point in this study may be the rubric itself. Many researchers share the advantageous contribution of rubric usage to the assessment and evaluation process. It helps both teachers and students have a common criterion for learning goals and outcomes. For teachers, rubrics provide an organized assessment criterion, and for students they provide a detailed feedback for the analysis of the weak and strong points of their learning. Stevens and Levi (2005) state on these issues as:

> Rubrics save time, provide timely, meaningful feedback for students, and have the potential to become an effective part of the teaching and learning process. There are many reasons to use rubrics, reasons having to do not only with efficient use of time and sound pedagogy but, moreover, with basic principles of equity and fairness (p. 17).

5.3. Limitations of the Study

In this study, the entire C and D level students' essays, 251 essays at total, were used while analyzing the present rubrics in terms of inter-rater reliability. However, as the time was limited and it was an analysis trial, 59 C and D level students' essays were analyzed for the inter-rater reliability of the newly developed rubric. In order to conduct a more extensive study and to get more reliable results, the number of graders and the student essays could have been increased for the newly developed rubric.

Another limitation of this study may be not analyzing the rubrics used for the whole levels from A (elementary) level to D (upper-intermediate) level at the institution. Only rubrics for C (intermediate) and D (upper-intermediate) levels were used for analysis. It was because only in those levels students are supposed to write essays, which caused this study to look at the essay scoring rubrics. If the time had been sufficient, the same procedures could have been applied to the rubrics used for level A (elementary) level and B (pre-intermediate) level whose main focuses are on paragraph writing.

The other limitation can be counted as; the newly developed rubric is institutional and has been developed in accordance with the writing class objectives of Zirve University English Preparatory School. Therefore, the rubric is context-specific. Whether it can be used and can work in similar contexts at other educational institutions are questionable.

5.4. Implications for Further Studies

In order to obtain more consistent reliability results in an institution, teachers or instructors using the rubric(s) should be trained in order to make them be familiar with the rubric in terms of applying and grading. While developing a context-specific rubric, a focus group can often meet, and individual ideas of each participant can be taken into consideration. A bottom-up rubric developing approach can be adopted, which strengthens the validity and makes the rubric(s) more reliable. During the rubric development process, teachers or instructors teaching and grading writings can have the opportunity to give constructive feedback to the rubric developer(s).

6. REFERENCES

Allen, S., & Knight, j. (2009). A Method for Collaboratively Developing and Validating a Rubric. *Vol. 3, No. 2*.

Andrade, H. G. (1997). *Understanding Rubrics.* Retrieved December 23, 2012, from http://learnweb.harvard.edu/ALPS/thinking/docs/rubricar.htm

Andrade, H. (2005). Teaching with Rubrics: The Good, The Bad, and The Ugly. *College Teaching , 53* (1), 27-30.

Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation , 10* (3), 1-11.

Bachman, L. (2004). *Statistical Analyses for Language Assessment.* Cambridge: Cambridge University Press.

Bachman, L., & Palmer, A. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests.* Hong Kong: Oxford University Press.

Becker, A. (2011). Examining Rubrics Used to Measure Writing Performance in U.S. Intensive English Programs. *The Catesol Journal , 22* (1), 113-130.

Bresciani, M. J., Oakleaf, M., Kolkhorst, F., Nebeker, C., Barlow, J., Duncan, K., et al. (2009). Examining Design and Inter-Rater Reliability of a Rubric Measuring Research Quality across Multiple Disciplines. *Practical Assessment, Research and Evaluation , 14* (12), 1-7.

Brown, G., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9*, 105-121.

Brown, J. D. (1996). *Testing in Language Programs.* New Jersey: Prentice Hall Regents.

Broad, B. (2003). *What We Really Value Beyond Rubrics in Teaching and Assessing Writing.* Logan, Utah: Utah State University Press.

Burke, K., Ouellette, J., Miller, W., Leise, C., & Utschig, T. (2012). Measuring Writing as a Representation of Disciplinary Knowledge. *International Journal of Process Education , 4* (1), 13-27.

Cho, K., Schunn, C., & Wilson, R. (2006). Validity and Reliability of Scaffolded Peer Assessment of Writing From Instructor and Student Perspectives. *Journal of Educational Psychology , 98* (4), 891-901.

Clinard, J. (2011, January). A Practical Guide to Writing Proficiency. *The Montana University System Writing Assessment* , 1-39.

Comer, K. (2011, December 22). *Developing Valid and Reliable Rubrics for Writing Assessment: Research and Practice.* Retrieved December 21, 2012, from www.akoaotearoa.ac.nz/gppg-ebook: http://akoaotearoa.ac.nz/ako-hub/good-practice-publication-grants-ebook/resources/pages/writing-assessment-research

Çetin, Y. (2011). Reliability of raters for writing assessment: analytic - holistic, analytic - analytic, holistic - holistic. *Mustafa Kemal University Journal of Social Sciences Institute, 8* (16), 471-486.

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book.* London and New York: Routledge: Taylor and Francis Group.

Grawe, N. D., Lutsky, N. S., & Tassava, C. J. (2010). A Rubric for Assessing Quantitative Reasoning in Written Arguments. *Numeracy , 3* (1).

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll, *Exploring the Dynamics of Second Language Writing* (pp. 162-210). New York.

Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's Correlation Coefficient on the Same Sets of Data. *Quaestiones Geographicae , 30* (2), 87-93.

Heaton, J. (2003). *Writing English Language Tests, New Edition.* Longman Pub Group.

Jonsson, A., & Svingby, G. (2007). *The use of scoring rubrics: Reliability, validity and educational consequences.* Educational Research Review , 2, 130-144.

Johnson, K., & Vosmik, J. (2007). Assessing Student Learning: A Collection of Evaluation Tools . *OTRP Online Office of Teaching Resources in Psychology* , 1-31.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? . *Assessing Writing , 16*, 81-96.

Kopriva, R. (2008). *Improving Testing for English language Learners.* New York and London: Routledge: Taylor and Francis Group.

Kurt, A., & İzmirli, S. (2009). *The views of teacher candidates about the use of a scoring rubric for the evaluation of their products in the course of instructional technologies and material development*. Procedia Social and Behavioral Sciences , 1, 988-992.

Lovorn, M., & Rezaei, A. (2011). Assessing the Assessment: Rubrics Training for Pre-service and New In-service Teachers. *Practical Assessment, Research and Evaluation , 16* (16), 1-18.

Mcnamara, T. (2000). *Language Testing.* London: Oxford University Press.

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

Mertler, C. (2001). Designing Scoring Rubrics for Your Classroom. *Practical Assessment, Research & Evaluation , 7* (25).

Mohan, K. M., Miller, J. M., Dobson, V., Harvey, E. M., & Sherrill, D. L. (2000). Inter-rater and intra-rater reliability in the interpretation of native American preschool children. *Optometry and Vision Science* , 77(09), 473-482.

Moskal, B., & Leydens, J. (2000). Scoring Rubric Development: Validity and Reliability. *Practical Assessment, Research & Evaluation , 7* (10).

Munoz, D. (2009). Reliability as a Context-Dependent Requirement for Writing Proficiency Assessment. *University of Reading: Language Studies Working Papers , 1*, 46-54.

Nutter, F. W., Jr., Gleason, M. L., Jenco, J. H., & Christians, N. C. (1993). Assessing the accuracy, intra-rater repeatability, and inter-rater reliability of disease assessment systems. *Phytopathology* , 83, 806-812.

Payne, D. (2003). *Applied Educational Assessment* (2nd Edition ed.). Belmont, CA: Wadsworth/Thomson Learning.

Peat, B. (2006). Integrating writing and research skills: Development and testing of a rubric to measure student outcomes. *Journal of Public Affairs Education , 12*, 295-311.

Pickett N., & Dodge B. (2007). *Rubrics for web lessons*. Retrieved March 22, 2013, from http://webquest.sdsu.edu/rubrics/weblessons.htm

Pitoniak, M., Young, J., Martiniello, M., King, T., Buteux, A., & Ginsburgh, M. (2009). Guidelines for the Assessment of English Language Learners. *Educational Testing Service* , 1-29.

Polly, C., Rahman, A., Rita, L., Yun, Y., & Ping, L. (2008). *An Investigation of Reliability and Validity: Using Rubric Approach in Learning & Teaching*. Retrieved February 10, 2013, from www.cambridgeAssessment.org.uk:http://www.cambridgeassessment.org.uk/ca/digitalAssets/151585_Lee_Yim_Ping.pdf

Rea-Dickins, P., & Germaine, K. (1993). *Language Teaching: A Scheme for Teaching Education: Evaluation*. Hong Kong: Oxford University Press.

Reineke , M. (2007). *Introduction to Rubrics.* the CHFA SOA Committee.

Reynolds-Keefer, L. (2010). Rubric-referenced assessment in teacher preparation: An opportunity to learn by using. *Practical Assessment, Research & Evaluation , 15* (8), 1-9.

Rudner, L., & Schafer, W. (2002). *What Teachers Need to Know About Assessment.* Washington D.C.: National Education Association.

Sezer, S. (2006). Öğrencinin akademik başarısının belirlenmesinde tamamlayıcı değerlendirme aracı olarak rubrik kullanımı üzerinde bir araştırma. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi , 18*.

Siddiek, A. (2010). The Impact of Test Content Validity on Language Teaching and Learning. *Canadian Center of Science and Education , 6* (12), 133-143.

Stellmack, M., Konheim-Kalkstein, Y., Manor, J., Massey, A., & Schmitz, J. (2009). An Assessment of Reliability and Validity of a Rubric for Grading APA-Style Introductions . *Teaching of Psychology , 36* (2), 102-107.

Stemler, S. (2004, January 3). *A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability.* Retrieved April 5, 2013, from pareonline.net: http://pareonline.net/getvn.asp?v=9&n=4

Stevens, D., & Levi, A. (2005). *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning.* Sterling, Virginia: Stylus Publishing, LLC.

Şahinkarakaş, Ş. (1993). *The Reliability of Holistic and Analytic Evaluations of EFL Essays by Turkish University Preparatory Students.* Ankara: The Faculty of Letters and Humanities and the Institute of Economics and Social Sciences of Bilkent University.

Tchudi, S. (1997). *Alternatives to Grading Student Writing.* Urbana, Illinois, The USA: the National Council of Teachers of English.

Tierney, R., & Simon, M. (2004, January 28). *What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels.* Retrieved March 30, 2013, from pareonline.net: http://PAREonline.net/getvn.asp?v=9&n=2

Turkkorur, A. (2005). *Wriitng Portfolio Assessment and Inter-rater reliability at Yıldız Technical University School of Foreign Languages Basic English Department.* Ankara, Turkey: Department of Teaching English as a Foreign Language Bilkent University.

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

Wiggins, G. (1998). *Educative Assessment. Designing Assessments To Inform and Improve Student Performance.* San Francisco, California: Jossey-Bass Publishers .

Yürekli, A., & Üstünoğlu, E. (2007). Towards Student Involvement in Essay Assessment. *Essays in Education , 22*, 55-64.

Zimmaro, D. (2004, January 13). *Developing Grading Rubrics.* Retrieved April 02, 2013, from http://www.utexas.edu/academic/mec:http://goglobal.fiu.edu/Faculty/Documents/Faculty%20Development%20Resources/ZIMMARO_Developing_Rubrics.pdf

7. APPENDICES

7.1. Appendix 1: C Level – The Present Essay Scoring Rubric

The first descriptor of the rubric ‘Holistic’ and its scoring values

0 points - There is nothing written

1 points - There are some phrases but no complete sentences.

3 points - The student lacks many writing skills. There are some sentences, but the composition is generally not coherent.

4 points - There are quite a few mistakes, and the student is on the border for failing.

5 points - The writing can be improved, but it is satisfactory for the level.

6 points - The writing is in a good condition. S/he has some errors.

8 points - The writing is in a very good condition with very few errors.

10 points - The student has written a well-developed composition for the level.

The second descriptor of the rubric ‘Title’ and its scoring values

0 points – No title

2 points – There is a title

The third descriptor of the rubric ‘Format’ and its scoring values

0 points - No Paragraph format

1 points - Less/more than three paragraphs (five paragraphs for D level)

2 points - Three paragraphs (five paragraphs for D level)

The forth descriptor of the rubric ‘Indent’ and its scoring values

0 points - There is no indent

1 points - There are less/more than 3 (in C level) / 5 (in D level) indents

2 points - There are 3 (in C level)/ 5 (in D level) indents

The fifth descriptor of the rubric 'Introduction Paragraph' and its scoring values

0 points - No introduction paragraph

3 points - There is an introduction paragraph with some problems

5 points - There is an introduction paragraph which is OK

8 points - There is a well-developed introduction paragraph

The sixth descriptor of the rubric 'Body' and its scoring values

0 points - No body paragraph

2 points - There is a body paragraph with many problems, lack of examples

4 points - There is a body paragraph with some missing details/examples

7 points - There is a body paragraph with little problems

10 points - There is a well developed body paragraph

The seventh descriptor of the rubric 'Conclusion Paragraph' and its scoring values

0 points - No conclusion

3 points - Bad conclusion

5 points - OK (not good, not bad)

8 points - Well developed-ending, conclusion

The eighth descriptor of the rubric 'Content' and its scoring values

0 points - Unable to develop a coherent idea, no sentences at all.

4 points - There is a composition, but not enough quality.

7 points - Satisfactory quality

10 points - Very satisfactory quality for the level

The ninth descriptor of the rubric 'Vocabulary' and its scoring values

0 points - Awkward sentences due to lack of vocabulary.

2 points - Some mistakes on the basic vocabulary they know. Errors in meaning.

5 points - Basic vocabulary only

8 points - Appropriate word choice but no surprises. Uses level vocabulary in appropriate places.

10 points - Creative use of new vocabulary for the level.

The tenth descriptor of the rubric 'Grammar' and its scoring values

0 points - No correct sentences

3 points - Many problems

6 points - OK but some problems

9 points - Good

12 points - Very good, nearly no mistakes

The eleventh descriptor of the rubric 'Punctuation' and its scoring values

0 points - No correct punctuation

1 points - Many problems

2 points - OK, but some problems

3 points - Good

The twelfth descriptor of the rubric 'Spelling' and its scoring values

0 points - No correct spelling

1 points - Many problems

2 points - OK, but some problems

3 points - Good

The thirteenth descriptor of the rubric 'Transitions/Connectors' and its scoring values

0 points - No use of transitions/connectors, or unnecessary usage.

2 points - Little use of transitions. Many problems with the meaning of the connectors / transitions.

5 points - Some use of transitions. Some problems.

8 points - Appropriate use of transitions. Few problems with the meaning.

10 points - Very good and sufficient use of transitions. Transitions are in appropriate place.

The fourteenth descriptor of the rubric 'Coherence' and its scoring values

0 points - No coherence/paragraph has more than one idea

5 points - Problems with the coherence or expressing one idea

10 points - Good coherence and one idea

7.2. Appendix 2: D Level – The Present Essay Scoring Rubric

The first descriptor of the rubric 'Holistic' and its scoring values

0 points - There is nothing written

1 points - There are some phrases but no complete sentences.

3 points - The student lacks many writing skills. There are some sentences, but the composition is generally not coherent.

4 points - There are quite a few mistakes, and the student is on the border for failing.

6 points - The writing can be improved, but it is satisfactory for the level.

8 points - The writing is in a good condition. S/he has some errors.

10 points - The writing is in a very good condition with very few errors.

12 points - The student has written a well-developed composition for the level.

The second descriptor of the rubric 'Title' and its scoring values

0 points – No title

1 points – There is a title

The third descriptor of the rubric 'Format' and its scoring values

0 points - No Paragraph format

1 points - Less/more than five paragraphs

3 points - Five paragraphs

The fourth descriptor of the rubric 'Introduction Paragraph' and its scoring values

0 points - No introduction paragraph

3 points - There is an introduction paragraph with some problems

5 points - There is an introduction paragraph which is OK

8 points - There is a well-developed introduction paragraph

The fifth descriptor of the rubric ‘Body’ and its scoring values

0 points - No body paragraph

2 points - There is a body paragraph with many problems, lack of details

4 points - There is a body paragraph with some missing details

7 points - There is a body paragraph with little problems

10 points - There is a well developed body paragraphs

The sixth descriptor of the rubric ‘Conclusion Paragraph’ and its scoring values

0 points - No conclusion

3 points - Bad conclusion

5 points - OK (not good, not bad)

8 points - Well developed-ending, conclusion

The seventh descriptor of the rubric ‘Content’ and its scoring values

0 points - Unable to develop a coherent idea, no sentences at all.

4 points - There is a composition, but not enough quality.

7 points - Satisfactory quality

10 points - Very satisfactory quality for the level

The eighth descriptor of the rubric ‘Vocabulary’ and its scoring values

0 points - Awkward sentences due to lack of vocabulary.

2 points - Some mistakes on the basic vocabulary they know/problems with the meaning.

4 points - Basic vocabulary only

6 points - Appropriate word choice but no surprises. Uses level vocabulary in appropriate places.

10 points - Creative use of new vocabulary for the level.

The ninth descriptor of the rubric 'Grammar' and its scoring values

0 points - No correct sentences

3 points - Many problems

7 points - OK but some problems

9 points - Good

12 points - Very good, nearly no mistakes

The tenth descriptor of the rubric 'Punctuation' and its scoring values

0 points - No correct punctuation

1 points - Many problems

2 points - OK, but some problems

3 points - Good

The eleventh descriptor of the rubric 'Spelling' and its scoring values

0 points - No correct spelling

1 points - Many problems

2 points - OK, but some problems

3 points - Good

The twelfth descriptor of the rubric 'Transitions/Connectors' and its scoring values

0 points - No use of transitions/connectors, or unnecessary usage.

3 points - Little use of transitions. Many problems with the meaning of the connectors / transitions.

6 points - Some use of transitions. Some problems.

8 points - Appropriate use of transitions. Few problems with the meaning.

10 points - Very good and sufficient use of transitions. Transitions are in appropriate place.

The thirteenth descriptor of the rubric ‘Coherence’ and its scoring values

0 points - No coherence/paragraph has more than one idea

5 points - Problems with the coherence or expressing one idea

10 points - Good coherence and one idea

7.3. Appendix 3: The Newly Developed Essay Scoring Rubric

The first descriptor of the rubric 'Essay Format (paragraphs/indents/title)' and its scoring values

0 points - Has no essay format.

3 points - Does not have the enough number.

5 points - Has the right number of the items above.

The second descriptor of the rubric 'Content (content/coherence)' and its scoring values

0 points - There is nothing written.

10 points - Content has serious problems. Ideas are not coherent.

20 points - Content is almost appropriate to the topic. Ideas may/may not be coherent.

30 points - Content is appropriate to the topic. Ideas are coherent

The third descriptor of the rubric 'Organization (introduction/body/conclusion)' and its scoring values

0 points - There is nothing organized.

5 points - Has so many problems with the thesis statement, topic sentences, or supporting details.

15 points - Has slight problems with the thesis statement, topic sentences, or supporting details.

25 points - Has a clearly stated thesis statement, topic sentences, and supporting details.

The fourth descriptor of the rubric 'Use of Language (grammar/spelling/punctuation)' and its scoring values

0 points - Has nothing correct.

5 points - Has so many problems and interferes with the meaning.

15 points - Uses language efficiently with slight mistakes and still comprehensible.

25 points - Uses language efficiently with almost/no mistakes.

The fifth descriptor of the rubric ‘Vocabulary (transitions/connectors/vocabulary)’ and its scoring values

0 points - Has nothing correct.

5 points - Not enough use.

10 points - Has the level-specified use.

15 points - Has a creative and correct use.