HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Department of Foreign Language Education
English Language Teaching Program

AN ANALYTIC APPROACH TO ENGLISH LANGUAGE INSTRUCTORS' SCORING DIFFERENCES OF WRITING EXAMS

Fatma Merve UZUN

Master's Thesis

Ankara, (2019)

With leadership, research, innovation, high quality education and change,

To the leading edge... Toward being the best...

HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Department of Foreign Language Education
English Language Teaching Program

AN ANALYTIC APPROACH TO ENGLISH LANGUAGE INSTRUCTORS' SCORING DIFFERENCES OF WRITING EXAMS

İNGİLİZCE OKUTMANLARININ YAZMA SINAVLARINDAKİ NOTLANDIRMA FARKLILIKLARINA ANALİTİK BİR YAKLAŞIM

Fatma Merve UZUN

Master's Thesis

Ankara, (2019)

Acceptance and Approval

To the Graduate School of Educational Sciences,

This thesis, prepared by **FATMA MERVE UZUN** and entitled “An Analytical Approach to English Language Instructors’ Scoring Differences of Writing Exams” has been approved as a thesis for the Degree of **Master** in the **Program of English Language and Teaching-Master of Sciences** in the **Department of Foreign Languages Education** by the members of the Examining Committee.

Chair	Prof. Dr. Nuray ALAGÖZLÜ	N. alagözlü
Member (Supervisor)	Assist. Prof. Dr. İsmail Fırat ALTAY	
Member	Assist. Prof. Dr. Ceyhun KARABIYIK	

This is to certify that this thesis has been approved by the aforementioned examining committee members on …/…/2019 in accordance with the relevant articles of the Rules and Regulations of Hacettepe University Graduate School of Educational Sciences, and was accepted as a **Master’s Thesis** in the **Program of English Language and Teaching** by the Board of Directors of the Graduate School of Educational Sciences on …../…../.2019

Prof. Dr. Ali Ekber ŞAHİN
Director of Graduate School of Educational Sciences

Abstract

This study investigated scoring results of writing exams assigned by raters so as to see how consistent and reliable scores can be acquired focusing on raters' individual scoring results and different raters' scoring results for similar written performances. Writing assessment is away from clear-cut answers, consequently, ensuring objectivity in scores has always been longed for. Various studies have been conducted to achieve this. The current study primarily aimed at discovering all factors affecting the scoring process in order to avoid variance in scoring results. For this, in a state university in Ankara, using 3 techniques- questionnaire, think aloud and interview-, data was collected and the research was conducted with 15 ELT instructors teaching at the School of Foreign Languages chosen by convenience sampling and 25 ELT instructors participating in the questionnaire. A mixed method research design was employed. For the quantitative findings, SPSS Kruskal Wallis and Mann-Whitney U tests were used and insignificant results were gathered, whereas quantitative results provided by percentages were quite significant. As for the qualitative findings, the analysis clearly illustrated that there are factors mainly like rater effects, rubric use, scoring styles, prioritized and ignored criteria, experience in teaching, comparing performances, failing to adapt to the level to be considered, and institutional goals causing both inter-rater unreliability and intra-rater unreliability. To obtain consistent results, using rubrics with well-defined criteria and categories, consultation and feedback, standardization meetings, frequent workshops can be pursued in addition to benchmarking and multiple scoring and future ELT instructors can be guided accordingly.

Keywords: scoring writing performance, rater reliability, rubric use, assessment, writing

Öz

Bu çalışma benzer yazılı performanslar için değerlendiricilerin bireysel değerlendirme/puanlama sonuçları içinde ve birbirinden farklı değerlendiricilerin puanlama sonuçlarına odaklanarak ne kadar tutarlı ve güvenilir sonuçlar elde edilebileceğini anlamak maksadıyla değerlendiricilerce puanlandırılan yazma sınav sonuçlarını incelemiştir. Yazılı performans değerlendirmesi keskin ve net cevaplardan uzaktır, dolayısıyla değerlendirme sonuçlarında nesnelliği sağlamak her zaman aranan bir durumdur. Bunu elde edebilmek için birçok çalışma gerçekleştirilmiştir. Bu güncel çalışma değerlendirme/puanlama sonuç farklılıklarından kurtulabilmek için değerlendirme sürecini etkileyen bütün faktörleri bulmayı hedeflemiştir. Bu amaçla, Ankara'da bir devlet üniversitesinde, anket, sesli düşünme ve görüşme teknikleri kullanılarak veri toplanmıştır ve araştırma ankete katılan 25 katılımcı ile birlikte uygun örneklemeyle seçilen 15 Yabancı Diller Yüksek Okulunda İngiliz dili öğreten öğretim görevlisiyle gerçekleştirilmiştir. Çalışmada nitel ve nicel araştırma yönetimlerince oluşan karma yöntem kullanılmıştır. Nicel sonuçlar için sosyal bilimler istatistik programı kullanılarak Kruskal Wallis ve Mann Whitney U testleri uygulanmış ve sonuçlar değersiz kalmıştır, ancak yüzdelik sonuçlar oldukça anlamlı çıkmıştır. Nitel veri sonuçlarına göre incelemeler, değerlendirici etkisi, değerlendirme ölçeği kullanımı, değerlendirme yöntemleri, ön plandan tutulan ve önemsenmeyen kriterler, öğretim deneyimi, öğrenci performanslarının karşılaştırılması, değerlendirilecek performans düzeyine uyum zorluğu ve yönetimsel hedefler gibi öne çıkan birçok faktörden kaynaklı tek hakem güvenilirliğin ve hakemler arası güvenilirliğin sağlanamamasına neden olduğunu açıkça göstermiştir. Tutarlı sonuçlar elde edebilmek için örnek üzerinde karşılaştırmalı değerlendirme ve çoklu değerlendirmenin yanı sıra iyi tanımlanmış kriterlerin ve kategorilerin olduğu değerlendirme ölçeği kullanımı, danışma ve geri bildirim, standardizasyon toplantıları, sıkça çalıştay düzenleme gibi yollar izlenebilir ve geleceğin İngiliz dili öğretim görevlilerine bu yönde rehberlik sağlanabilir.

Anahtar Kelimeler: yazılı performans puanlandırılması, değerlendirici güvenilirliği, değerlendirme ölçeği kullanımı, değerlendirme, yazma

To my beloved family,
my dear husband and
my little chickpea

Acknowledgments

In the first place, I would like to thank my supervisor Assist. Prof. Dr. İsmail Fırat Altay not only for his constructive feedback and guidance throughout my research, but also his consistent encouragement to make me carry out this study and write my thesis when I felt hopeless and fatigued and I want to express my gratitude to him for accepting me as a student and his understanding after I had to change my supervisor because of his retirement. I also thank my previous supervisor Prof. Dr. İsmail Hakkı Mirici for his continuous support and care in addition to his guidance during my health problems and surgery. I would like to thank Assoc. Prof. Dr. Hüseyin Öz for his help while developing my questionnaire and encouragement to complete my study.

I would like to thank the committee members Prof Dr. Nuray Alagözlü and Assist. Prof. Dr. Ceyhun Karabıyık for their valuable comments and suggestions to my research.

My sincere appreciation is for my dear family who always care and encourage me to conduct my study and for their belief in me. My father Adem Ölmez, my mother Neriman Ölmez and my brother Mesut Ölmez have always provided me with their unconditional love and support throughout all my education life.

My deepest thanks go to my husband Yakup Uzun for his never ending confidence in me and feedback to my study.

I am appreciative of all the instructors participating in my study for their contributions and students allowing me to use their writing exams.

Table of Contents

List of Tables

List of Figures

Symbols and Abbreviations

CEFR: Common European Framework of References for Languages

ELT: English Language Teaching

EFL: English as a foreign language

ESL: English as a second language

FLE: Foreign Language Education

G: Teacher group made according to years of experience

L2: Second/Foreign language

P: Student writing exam

SPSS: Statistical Package for the Social Sciences

T: Instructor /Rater

Chapter 1
Introduction

Introduction

As is known, writing is a significant component of English language teaching and its testing and evaluation have always been a concern to teachers, students, administrators because of its quality of performance, so scholars have worked hard to find a solution to this issue. In this section, the problem which is the starting point of the study, some background information about it, the purpose of the study, the significance of the study, the research questions aimed to find out about the problem, the assumptions and expectations of its contributions to English teaching and testing, the obstacles encountered while conducting the study and definitions of terms regarding the study will be stated.

Statement of the Problem

Determining whether students pass or fail is one of the crucial judgments in education. To show their own developmental progress, students take various exams, some of which include performance assessment during a term, or a year. In foreign language education, assessing writing can be challenging as there comes out rater effects, which questions the reliability or validity of measure. Therefore, Cumming, Kantor and Powers (2002) stated that the need for learning how the raters decide in their final judgment for student writing is significant. To guide raters to decide on a single score, analytical and holistic assessments are widely used in writing. Instructors, as raters, are trained to grade student writing in the same way if they are given a previously set of criteria as a guidance (Vaughan, 1991). Hence, it is assumed that each rater will score student writing with little or no difference when given a scale of descriptors. It is surely beyond doubt that the scale or rubric is to be well developed. According to Eckes (2008), 'scoring criteria play a crucial role in rater-mediated performance assessments' (p.156). With the help of the criteria, instructors assess students' writing, come to an opinion about them, and score the exams of students. As all raters use the same scoring criteria, it should be tested in terms of validity and reliability beforehand. However, no matter how it is tested, instructors may focus their attention on a particular element of student writing such as vocabulary, content, organization, grammar, etc.

A great many of components are regarded while scoring writing: length of writing, spelling, punctuation, grammatical structures, how communicative it is, content, lexicon, semantics, clearness, organization, etc. (Milanovic, Savielle &Shuhong, 1996). Raters may get influenced more by one of these elements, which leads to rater effects in scores. On a large scale rubric, White (1985) stated that 1-point scoring difference is acceptable, whereas the difference more than 3 means a big problem as it may cause student failure in some situations. Instructors' scoring styles have a critical role in writing assessment, therefore, Vaughan (1991) points out 'it is also important to evaluate the process by which raters make their decisions' (p.111). How instructors judge student writing should be explained precisely for the sake of reliability.

Background of the Study

While scoring writing exams, closing the grading gap among the instructors has been longed for in foreign language education field, but it is really demanding to find out about the underlying reasons as in every foreign language school or institution, the practices of scoring writing may change. To give an example, using ready-made scoring rubrics, to say, just picking up a rubric randomly from the internet or adapting it without enough analysis of students' needs is a reason for scoring differences (Tierney & Simon, 2004). If instructors have a grasp of the elements of scoring rubric given to them and do enough practice with them, the difference may decrease, even disappear.

Raters attaching importance to different categories on a large scale scoring rubric cause inconsistency among them. The variability of writing scores may result from many elements such as how much instructors contribute to the scoring rubric, how effectively they use it, how they interpret the components in the rubric, the complexity or the inadequacy of the rubric, how consistent the rubric is when used for different levels of students, whether the raters are trained enough to use the rubric efficiently, how experienced the raters are in terms of scoring writing and raters' background. No matter what purpose it is used for or it is designed or adapted, the crucial point is its being consistent (Tierney & Simon, 2004).

Different studies assert different reasons for scoring variety. Cumming (1990) stated that the raters who are more experienced prefer using a large scale rubric including details in contrast to less experienced instructors. Some raters may not need a scoring rubric to judge writing exams, they just assess student papers overall and decide if they are good or not, however other raters cannot evaluate the exams without guidance. Eckes (2008) argued that raters cannot focus their attention on each category on the scoring rubric equally. Instructors concentrate on different aspects such as grammatical structures or vocabulary items more as they each have different reading styles (Vaughan, 1991).

To find common ground, Common European Framework of References for Languages (CEFR) is a very important instrument for curriculum guidelines and language syllabuses. Six levels of foreign language proficiency are described, from A1 to C2, so that teaching and learning objectives, materials, curriculum, assessment, etc. can be designed or improved. There are also ‘plus levels’ for stronger language performances as A2+, B1+ and B2 + (North, 2005). These plus levels have both criteria of previous and following levels. Institutions can use CEFR and its descriptors to develop internationally accepted materials and examinations. Moreover, reliable exams and scoring criteria can be made in the light of CEFR as instructors know what to expect from their students and what students can do. As stated in CEFR self-assessment grid, a student knows what s/he can do for listening, reading, writing, spoken interaction and spoken production with the self-assessment grid. To illustrate, for writing, a B2 level student is expected to say:

‘I can write clear, detailed text on a wide range of subjects related to my interests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. I can write letters highlighting the personal significance of events and experiences’ (Council of Europe, 2001, p.27).

Considering this ‘can do’ statement, instructors who know that their students are B2 expect their students to write an essay or a paragraph giving enough details. In this sense, while grading student writing, as students’ levels and backgrounds are known, the possible scoring differences might raise a question mark in minds.

Purpose of the Study

This study aims to examine English language instructors' scoring differences of the writing exams and search for instructors' attitude and preferences or tendencies while grading student writing. Whether the scoring difference is because of the materials used while grading, or rater effects will be searched. The possible causes of variance in scoring proficiency writing exams because of raters' own assessment in their own way or different raters' disagreement are aimed to reveal. If so, what makes disagreements on the same writing exam and what leads a rater to assign different scores to similar writing exams will be investigated. Another prominent purpose of the study is to attain real thoughts of the raters in a safe environment so that the best results for differing scores can be learned and the main reasons for the problem can be designated. So as to find effective solutions and help the study reach its purpose, getting the most correct results is crucial. Finally, for the reliability and validity of measurement, terminating the possible reasons is what is expected.

Significance of the Study

A way of showing language skills in performance, speaking and writing skills necessitate objective assessment, but can be exposed to subjectivity especially during scoring, which can lead to problems especially when students are tested in these skills. These two skills are productive skills and performance assessment is used in their testing and evaluation. Since assessing writing is quite difficult as human factor is included in the process and it shows up subjectivity in scores, many schools and universities try to understand why their instructors think and judge the writing exams differently from each other, why the same rater's focus, consequently final marks, can change in similar exams in spite of knowing student backgrounds or having experience in writing assessments where they work. This subjectivity issue is away from being a problem in some institutions especially private schools as they can use technical support in assessment and exclude human factors. However, in many state schools and universities, using human raters is inevitable.

Scoring difference appears in different universities for different reasons although a lot of studies have been made to reveal possible reasons for scoring variance and how to lessen the difference in scoring students' exams. Different from other studies, this study is expected to contribute to schools with the same problem by providing an insight, comparing their situations with this study results and showing a reflection to them. As the target participants are young adults (aged 17-23) with different departments who study at a preparatory school to pass the proficiency exam (B1+ level- CEFR), the study will be a great source for instructors, curriculum members, testers, and administrators in many ways such as preparing programs, exams and training instructors etc. With the up-to-date findings presented in my research, the factors determining scoring difference in writing exams will be explained in detail so that the problem which may even affect students' success or failure is expected to be solved.

Research Questions

1. Is there a significant difference in EFL instructors' scoring results in terms of their years of experience?
2. Is there a significant difference in EFL instructors' scoring results in terms of
   a. their use of rubric,
   b. their familiarity with the rubric
   c. their holistic and analytic scoring preference
3. Do the instructors have difficulty in scoring students' writing? If yes, what kinds of difficulties do they have?
4. Is there a significant difference between/among the student grades when their writing exams are scored by more than one instructor?
5. While grading students' writing, do the instructors pay their attention to a particular criterion more than others? If so, which criterion or criteria do they focus on more, and why?

Assumptions

It is assumed that the findings and solutions of this study will contribute to English Language Teaching (ELT), foreign language education (FLE) and assessment considering different practices of writing exams and scoring processes in different schools.

When a writing exam is scored by more than one rater with or without using a scoring rubric or checklist, raters' marks may not match with each other, which causes not only time wasting as the exams with variant scores need to be rescored by another rater or rater committee, but also unfairness in scores as a writing exam can get a higher or lower score than it is worth just because raters having similar tendencies or similar focuses on assessments can be chosen incidentally. This can result in assigning a score that doesn't reveal the exact level of the students, which harms reliable and valid assessment.

Similarly, for similar writing exams, if a rater assigns differing scores to similar writing exams and these variant scores make a difference in final scores, this inner inconsistency also causes serious problems for students and institutions.

In the light of all these issues, with this study, it is believed that considering how common this situation is in many schools and universities, the results of the study and pedagogical implications will be made use of in English language teaching, testing and evaluation.

Limitations of the Study

As in every research study, this study also encountered some limitations and unexpected situations and these limitations will be explained below.

In the School of Foreign Languages I carried out my research, the study was conducted with 15 English language instructors who were divided into 3 groups according to their years of experience. There was no instructor with less than 3-year experience and there were only 5 instructors whose experience is less than 10 years, moreover, not all of them were volunteer to participate in the study, which made it hard to find enough instructors for 'less experienced group'. Additionally, since the study included 3 research tools- questionnaire, think-aloud, interview, and it would take the participants' time to attend them all, the number of volunteer instructors was limited with 15 and the number of writing exams to be scores was limited with 12. If there were more writing exams to be scored, or the number of raters who could participate in the study were higher, the finding of the study could be different especially quantitatively. Finally, the writing rubric the raters used in assessing was updated before my application, so the new rubric might be new to the instructors although it was used in more than 5 exams.

Definitions of Terms

Scoring: It refers to giving a mark to students' performance.

Scoring writing: It refers to foreign language instructors' grading students' writing exams based on previously defined scoring criteria provided with them.

Assessment: It refers to estimating the quality or value of something or someone

Evaluation: It refers to judging the quality of student performance.

Evaluation checklist: It refers to a list of items showing the things to be considered in evaluating writing exams as a reminder.

Rating scale: It refers to a range of values used to judge and grade writing exams.

Writing rubric: It refers to a checklist that includes criteria questioning how adequate students are in grammar, vocabulary, organization, fluency, spelling, punctuation, tone, etc.

Rater: It refers to a person or a tool that judges something according to some standards or a scale and assigns a value.

Holistic evaluation: It refers to assessing students' writing as a whole and making a judgment according to a scale given to them before.

Analytical evaluation: It refers to assessing students' writing analytically, breaking the student writing into smaller pieces and evaluating each piece in detail according to a scale given to them before.

The consistency between raters: It refers to how similar or different the raters' grades are while scoring students' writing.

Inner consistency of a rater: It refers to how similar scores the same rater assigns to similar students' writing exams.

Conclusion

Writing assessment is one of the most critical type of assessments and as it doesn't have concrete answers to be provided, it is prone to subjectivity. This subjectivity can stem from factors like raters' personal ideas, focuses and considerations of each aspect of writing exams, their experience in scoring, etc. Not having one single and unchanged answer, students' writing performances can get variant scores when they are assessed by different raters, which causes unreliable scoring results and can affect students' success and failure. The variance may also come out due to the same rater's own scoring difference for similar writing performances. In order to find out the reasons behind such variance in scoring and to find applicable solutions, the study has been conducted and its findings are believed to contribute to any institution, school or university where writing assessment is conducted and to all further studies which will be realized about this issue.

Chapter 2
Literature Review

Introduction

On writing assessment and obtaining objectivity, reliability and validity in scores, different studies have been conducted. For this current study, in the light of these studies, some prior knowledge on what assessment is, the importance of reliable scoring and rater consistency in assessing writing, the role of rubric use in assessments are stated in the Background Information section. Details about objective assessment in writing, factors affecting the assessment process and results are stated in Assessing Writing section. The effect of rubric use in writing assessment, raters' use of rubrics, what is expected from a B1(+) student according to CEFR and what students can do in writing considering CEFR descriptors are explained in Scoring Methods section. What is holistic scoring and what is holistic rubric, what effects it has on reliable assessment are stated in Holistic Rubric section. What is analytical scoring and analytical rubric, the role of analytical scoring in writing assessment are given in Analytical Rubric section. Raters' effects as human factors on scoring writing are clarified in Rater Consistency section. Different raters' agreement or disagreement on writing assessment and how such disagreement influences will be stated in Inter-rater Reliability section and raters' own consistency in judging writing exams and the reasons behind possible inconsistency and its effect on writing scores are stated in Intra-rater Reliability section.

Background Information

Language learning assessment at all levels for all skills is a really important part of the learning process. Among the five skills (listening, reading, writing, spoken interaction and spoken production), the productive skills are more difficult to be assessed because scorers, usually human scorers, are required to judge the examinations, which could not make the assessment objective anymore. However, assessment has to be valid and reliable, free from who carried out it and how s/he scored, it should give the same results (Jonsson & Swingby, 2007; Moscal & Leydens, 2000; North, 2007). Since writing assessment necessitates higher order thinking process and cognitive skills, these can only be achieved by open ended tasks (Jonsson & Swingby, 2007). Different from multiple choice questions or responses involving objective scores, assessing writing as a productive skill is more demanding because writing prompts, scoring rubrics and scorers need to be taken into consideration (Siegert & Guo, 2009). In case the test score is reached subjectively, it is crucial to reduce or do away with rater inconsistency (Yen, 2016).

Assessment will be more reliable as long as a scoring rubric is used as it is a guide to help teachers as assessors score students' writings more objectively. Nothing has been observed to decrease the reliability of assessment when a rubric is used and consequently it will increase objectivity (Jonsson & Svingby, 2007; Silvestri & Oescher, 2006; Spandel, 2006). However, it is a fact that what rubric is used - holistic or analytic- ought to be carefully chosen in accordance with the purpose of exams (Bacha, 2001). Common European Framework of References (CEFR) is used for 'what is assessed, how performance is interpreted and how comparisons can be made' (Council of Europe, 2001, p.178). For assessment, descriptors can be applied as a scale in which various descriptors come together to direct assessors to make decisions on paragraphs in different levels. A kind of standardized criteria may help scoring process become easier.

To see how reliable a writing assessment is, any factors that affect test-takers' ultimate scores and assessors' scores should be paid attention to. Whether there is significant difference among raters while scoring students' writing and if such a difference is discovered and this difference is huge enough to change the score results of students and thereby influence the final score for a fail/ pass or correct level placement, this needs to be searched in detail (Meier, 2012).

So as to understand instructors' scoring differences of writing exams, it is necessary to observe how writing exams are assessed, which scoring method, holistic or analytic scoring, can be more effective depending on the education settings, how consistent raters can be while scoring writing and whether CEFR descriptors can contribute to hinder scoring differences.

Assessing Writing

Writing is a very good way to check how students can develop critical thinking skills, effective communication, creative learning, defending an argument and measure comprehension of content, so it is very important to increase the awareness of students for developing writing skill. Students' writing is assessed through paragraphs and essays, which is the most common method used in education. Writing skill is assessed in different ways such as for proficiency assessment, achievement assessment, performance assessment, self assessment, etc. (Council of Europe, 2001).

Scoring writing necessitates making decisions and concrete judgments about each student's language competence (Barrett, 2001). Some assessments are more critical for students. In a performance assessment, students are just required to give a sample of what they learned in a test showing casual progress, whereas students need to show how much they understand at the end of the week or terms in an achievement assessment, or they show their potential of what they can do in a proficiency assessment, which result in a pass/fail. Therefore, it is essential to analyze how assessment is performed. Çetin (2011) states that it is really challenging for teachers to score students' writing no matter how- holistically or analytically- as they encounter many problems like limited timing, trying to ensure objectivity in scores, finding fair writing prompts to consider and give a score. Among these problems, subjectivity is the most troublesome, as a matter of fact there is a variety of research on how to cope with subjectivity (Attali, 2016; Alanen, Huhta & Tarnanen, 2010; Bachman, 1990; Çetin, 2011; Davidson, Howell & Hoekema, 2000; Eckes, 2005; Fisher, Brooks & Lewis, 2002; Kayapınar, 2014; Kondo-Brown, 2002; Meier, 2012; Moscal, 2000; Schaefer, 2008; Spandel, 2006; Vaughan, 1991)

Subjectivity in performance testing cannot be avoided inevitably. In large scale writing assessments, the reliability is lower compared to multiple choice tests especially when all these exams take place under a time limit (Attali, 2016). Writing assessment cannot be conducted with clear-cut answer key as there is not only one correct answer in performance assessments. Unlike direct assessments, numerous factors are included in scoring. As the purpose is just to see student ability in how to communicate in English as a foreign or second language (EFL, ESL), avoiding rater biases or any irrelevances or inconsistency hindering what should be actually assessed is what is needed. (Schaefer, 2008). The point is how raters behave, think and the things they give special attention to while scoring writing is an issue of concern. (Attali, 2016; Bachman, Lynch, & Mason, 1995; Daly and Dickson-Markman, 1982; Eckes, 2008; Engelhard, 1994; Engelhard & Myford, 2003; Hughes, Keeling & Tuck, 1980; Lumley & McNamara, 1995; Schaefer, 2008; Weigle, 1998; Weigle, 1999).

Many studies focus on rater effects- how they differ from each other while interpreting scoring criteria. Among the common rater effects, overall reading, being lenient or severe to students' performances, being unable to give a final score considering different categories on the rubric, which is called 'halo effect', showing hesitation on scoring papers of the students who have the best or worst performances are found out (Engelhard, 1994; Meier, 2012). On this issue, Attali (2016) states that it is necessary to make sure all concerned raters find common ground on what composes qualified student writing in order to achieve consistency. Some researchers are also questioning the role of teaching or rating experience in the emergent differences of scoring (Pula & Huot, 1993; Wolfe, 1997; Wolfe, Song & Jiao 2016). They tried to figure out to what extent raters' experience can influence the rating process and the result. Weigle (1999) claimed that teachers without enough experience on scoring writing showed severity considerably compared to more experienced ones. Inexperienced scorers were inconsistent and more lenient in their own ratings, which improved with the help of training (Weigle, 1998). What is more, experienced raters are claimed to spend less time in determining their final score (Yen, 2016). They are quicker to make decisions because of the experience and they do not have any difficulty or much hesitation while rating. On the other hand, it is also claimed that no matter how much raters are trained, the differences in scores given by different raters can't be wiped out completely, let alone a meaningful reconciliation in the disagreement (Eckes, 2005).

Student writings are assessed using different scales mainly involving sections like content, organization, language use and accuracy, mechanics, vocabulary, flow of ideas, task response etc. Moscal (2000) argues that raters act differently and might pay attention to different properties of student writing like giving higher scores to those whose grammar structures are great or to those who are good at conveying his or her ideas in an effective way and convincing the readers of their arguments. While raters may be harsher to linguistic structures, they can be more lenient to organization (Schaefer, 2008). Or just the opposite can be observed. As Altay (2008) states, ‘without organization, all the valuable skills and components are wasted…idea generation, topic finding, narrowing down ideas, grouping them, eliminating irrelevant points and outline making are presented as main concerns of organization in writing’ (p.208). If students fail in narrowing down the ideas or mention irrelevant sentences, their score can be much lower than they expect just because they are not careful with organization. Moreover, raters who are more focused on organization may expect all sentence types for the flow of ideas as Demirezen (1993) states, simple, complex, compound and compound-complex sentences are main parts of organization, so students writing just simple sentences with correct use of grammar may get low grades because of this. Or, raters may not be sure about their judgments and give much lower or higher grades than deserved especially when the student shows a great or low language ability on his or her paper (Kondo-Brown, 2002). Meier (2012) calls this inconsistency ‘erratic rater behavior’ which means the unstable behaviors of raters (p.50).

The ways raters judge students’ performance can vary (Black, 2002). Scores cannot show parallelism to the similar papers although a specific rubric is used. Moreover, raters may not stay faithful to the scoring rubric unconsciously and add their personal opinion (Moscal, 2003). Or, without considering the descriptions and instructions on the rubric, raters score students’ writings overall, depending on their personal considerations (Çetin, 2011). To cope with this difference, double-marking has been used as an option. In a study by Breland, Camp, Jones, Marris and Rock (1987), the reliability when student writing is scored by one rater is lower than when it is scored by two raters. According to Yen (2016), ‘To ensure reliability in evaluation, every writing paper should be marked by at least two different raters. Each rater will mark independently. The score that the candidate receives for a piece of writing is the mean of the scores given by the two raters’ (p.81). In this way, while a rater may incline to focus on the linguistic features more and score student writing exams accordingly, another rater can tend to look for how fluently students can express their ideas and how clear they can give the message in their writing papers. These different viewpoints, conscious or unconscious subjectivity, divergent attitudes and personal interpretations are considered to be hindered by including more than one rater into the scoring process. However, there needs to be more research that aims at finding out the real reasons for instructors' scoring differences of writing exams, what affects the scoring differences most and whether these differences are changeable for the sake of attaining reliability.

Another significant factor in assessment is feasibility. In performance assessments, practicality of assessment is required. (Council of Europe, 2001). However, while assessing writing, under some conditions like time limit, lack of benchmarking to see various samples as much as possible before rating or lack of explanations or branches in the criteria they use, raters cannot be practical enough. This situation may also affect the scoring rubric choice. (Altay, 2007; Brown, 1994; Heaton, 1975).

Scoring Methods

In writing assessment, judging how well students can communicate and scoring their performance objectively are quite crucial. Using a scoring rubric is a powerful way of gaining a greater understanding of how proficient a student is in language learning (East & Cushing, 2016). Among the skills, writing helps understand a learner's performance in the target language, but as in writing assessment, performance assessments are more open to subjectivity and this questions reliability (Fisher et al, 2002). To deal with this, using an already developed scoring criteria can help (Çetin, 2011; Jonsson & Svingby, 2007; Moscal, 2000; Spandel, 2006). Since raters have a guide in hand, they cannot be a law unto themselves, or they may not have a difficulty making decisions on the final score when they use a rubric. Rubrics help and support raters in many ways, which directly affects its being applicable to use in any educational settings and demolishes any reliability concerns (Yuan & Recker, 2015). If there is not enough leading, raters can be out of the way. Actually, It has been already accepted that using rubrics increases the quality of assessment (Jonsson & Svingby, 2007). As they are adaptable and changeable depending on the purpose of the assessment, their use is extensive, from large scale assessments to personal assessment, so they have more benefits than drawbacks (Tierney & Simon, 2004). Rubric is one of the two elements of performance assessment (Perlman, 2003), therefore it needs a careful scheme and it must be understandable. It shouldn't be open to interpretation. According to Stemler (2004), there are three approaches that show accurate and consistent scoring: consistency estimates, consensus estimates and measurement estimates. The equivalence of scores between different raters, the tendency to give similar scores to the similar performances, and managing error-free scoring are what is desired in a good assessment.

Rubrics are used to assess different student tasks to show their levels in performance. (Hafner & Hafner, 2003). It can be used for any purposes: to assess basic or complex student work showing how much students can meet the criteria, to say, how much learning is achieved can be understood with an educational rubric (Jonsson & Svingby, 2007). Sometimes raters have trouble in managing with large number of students especially in high stake exams and they may assess students writing haphazardly, not sticking to a specific rubric. This may result in variation, moreover, because of not obeying rubric criteria and assigning a higher or lower score to the same performances can lead to success or failure. Supposing students want to understand the weak and strong points in their performance after seeing their mark, they cannot be provided so long as scoring rubrics aren't used. Rubrics give descriptions of what is expected for each level and they are useful for both guiding raters in giving marks to students and illustrating the problems in students' performances so that they can improve learning (Moscal, 2000).

Rubrics guide not only teachers but also students in both learning process and the output of this process (Andrade, 1997; Brookhart, 1999). They give feedback to students' performances, acknowledge students about what teachers expect from them and make students responsible for their own learning as they are the first judge of their own performance, which also promotes individual learning and self assessment. Additionally, when teachers use rubrics to assess writing, it becomes easy to explain why they give that score to the student and students can understand their own weaknesses and strengths (Andrade, 1997).

Writing assessment is not like scoring multiple-choice questions in which the answers are sole. 'In direct performance assessment, grades are generally awarded on the basis of a judgment. That means that the decision as to how well the learner performs is made subjectively, taking relevant factors into account and referring to any guidelines or criteria and experience' (Council of Europe, 2001, p.188). To deal with subjectivity, it is necessary to differentiate scoring writing with a rubric and by impression. When a rubric is used, student scores are based on a scale involving different bands to show the level, whereas when student performance is assessed by impression, no criteria is used as a reference to show the level, which causes a totally subjective, unproven results (Council of Europe, 2001). Therefore, it is better to use a rubric than not use it in order to avoid full subjectivity and overall impression (Spandel, 2006).

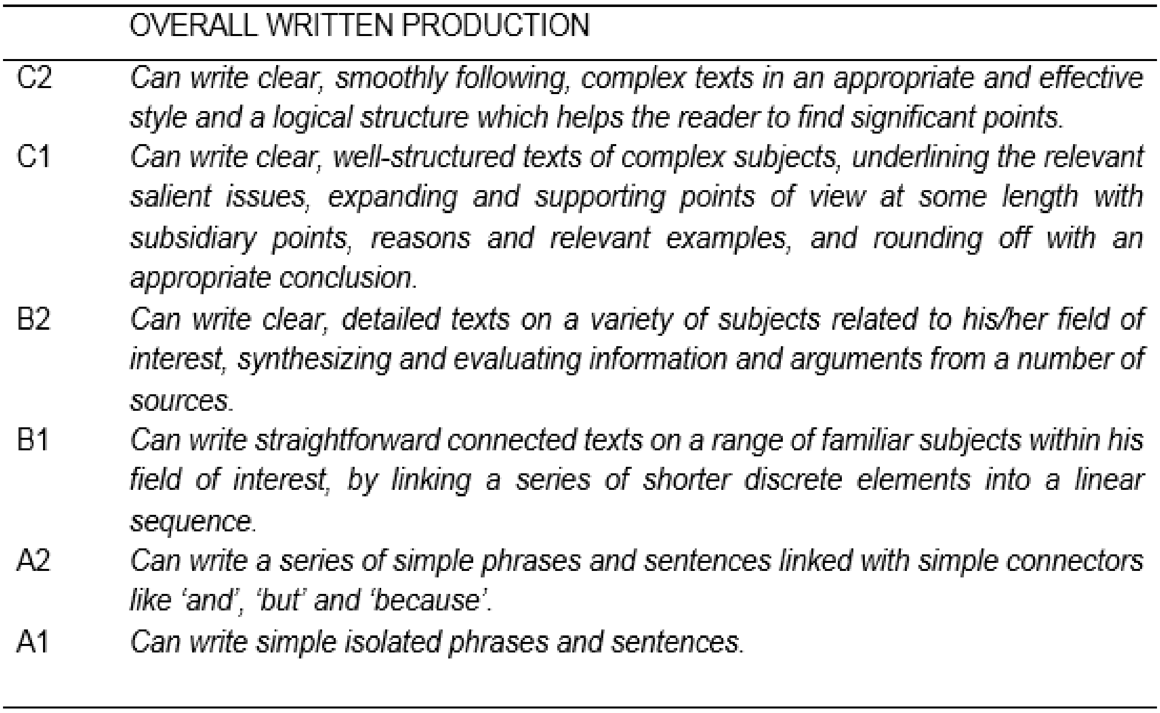
Using a rubric is advantageous in that it helps to assure consistency in scores and validity of judgments in performance assessments that are potentially prone to subjectivity (Jonsson & Svingby, 2007). However, there are also ideas about the inefficiency of using rubrics. In some studies like the one conducted by Rezaei and Lavorn (2010), it’s been claimed that raters who used a rubric couldn’t give enough attention to every aspect of the criteria, underestimate the content of writing and focused more on mechanics on student papers. No matter how valid and reliable the rubric is, scoring results cannot be accurate unless raters use the rubric effectively and sufficient piloting in using rubrics has been made. Meier (2012) states that rubrics aren’t functional and efficient as they are believed since they include many details that can cause confusion among raters. It would be better if they were made more applicable. What is more, we shouldn’t fail to notice the necessity of using scoring rubrics, yet they are not adequate alone for full reliability and accuracy on the assessment (Alanen et al, 2010). As long as the features of each scoring band that describe performances change for different levels and this causes inconsistency, neither the assessment can be valid nor learners can realize their mistakes (Tierney & Simon, 2004).

Rater training on the use of rubrics, sample selection of student performance to come to an agreement and careful conduction of the whole rating process are as important as rubric selection. Raters’ not using a specific rubric designed after careful considerations can also be an obstacle to a good judgment because standard rubrics cannot suit for each exam type although teachers might have to use them from time to time. Therefore, choosing the most reliable type of scoring rubric to test foreign language skills may result in a good deal of validity compared to other exam types involving test items (Wang, 2009)

Scoring rubrics are mainly used in performances assessments including communicative language activities like speaking and writing (Council of Europe, 2001; Moscal, 2000). As rubrics are illustrative scales giving descriptions of expected language skills in separate bands and they can be used not only in formal examinations, essay or paragraph writing, reports, but also in creative writing or in general writing. In CEFR, depending on the type of examinations, different scales showing all levels are suggested. As rubrics help learners to guide how and what to assess by providing clarifications and expectations, they lower the impact of variability in rating (Janssen, Meier & Trace, 2015)

Table 1

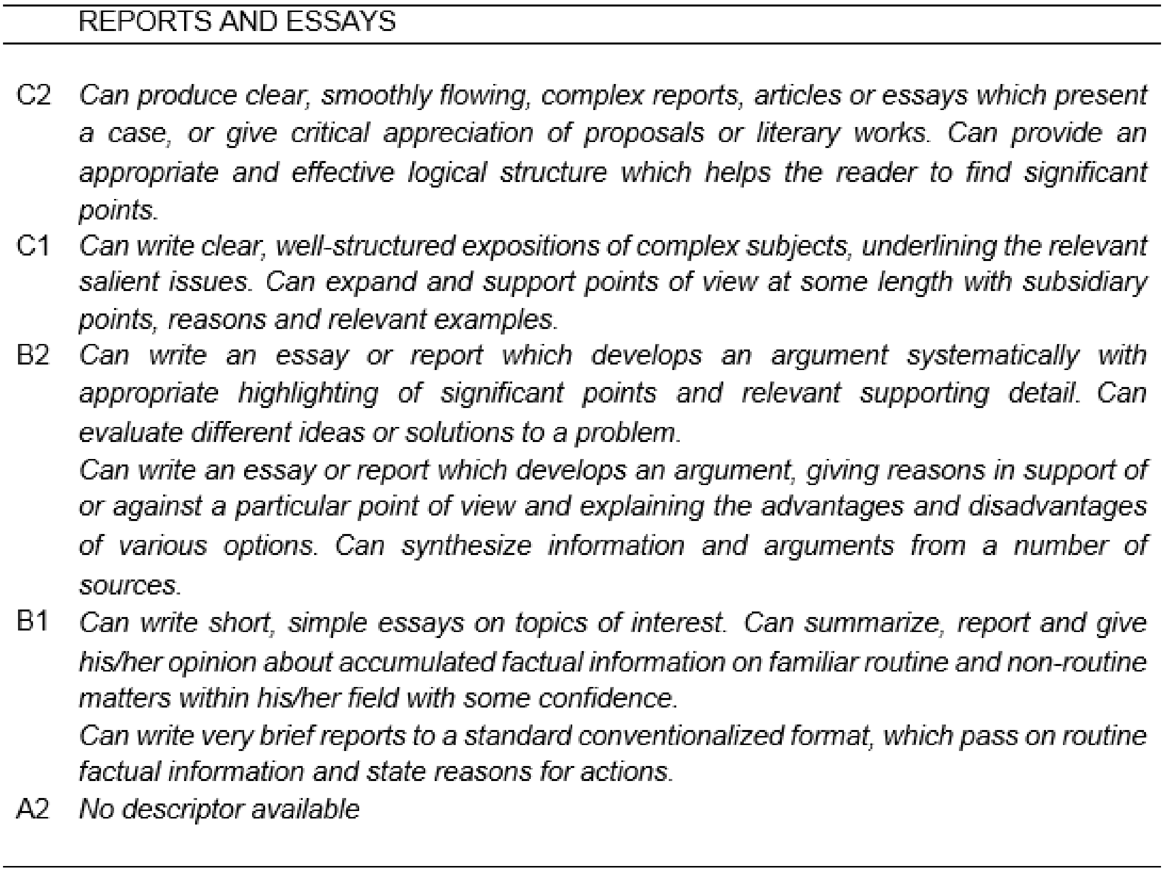
CEFR Overall Written Production - Illustrative Scale



	OVERALL WRITTEN PRODUCTION
C2	*Can write clear, smoothly following, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points.*
C1	*Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion.*
B2	*Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesizing and evaluating information and arguments from a number of sources.*
B1	*Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence.*
A2	*Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.*
A1	*Can write simple isolated phrases and sentences.*

Table 2

CEFR Reports and Essays-Illustrative sub-scale



	REPORTS AND ESSAYS
C2	*Can produce clear, smoothly flowing, complex reports, articles or essays which present a case, or give critical appreciation of proposals or literary works. Can provide an appropriate and effective logical structure which helps the reader to find significant points.*
C1	*Can write clear, well-structured expositions of complex subjects, underlining the relevant salient issues. Can expand and support points of view at some length with subsidiary points, reasons and relevant examples.*
B2	*Can write an essay or report which develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail. Can evaluate different ideas or solutions to a problem.* *Can write an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. Can synthesize information and arguments from a number of sources.*
B1	*Can write short, simple essays on topics of interest. Can summarize, report and give his/her opinion about accumulated factual information on familiar routine and non-routine matters within his/her field with some confidence.* *Can write very brief reports to a standard conventionalized format, which pass on routine factual information and state reasons for actions.*
A2	*No descriptor available*

The tables taken from CEFR (Council of Europe, 2001, pp. 61-62) above just display what a learner can do in each level to show their linguistic performance in writing. If B1 is taken as an example, a student of B1 level is expected to write an essay briefly, state his/ her opinions on a given topic by giving reasons and write reports. When the level is higher than B1 and lower than B2 at the same time- B1+ level, a student can even write an argumentative essay or benefit and drawback essay with stating examples and explanations. In terms of creative thinking, a B1 level student:

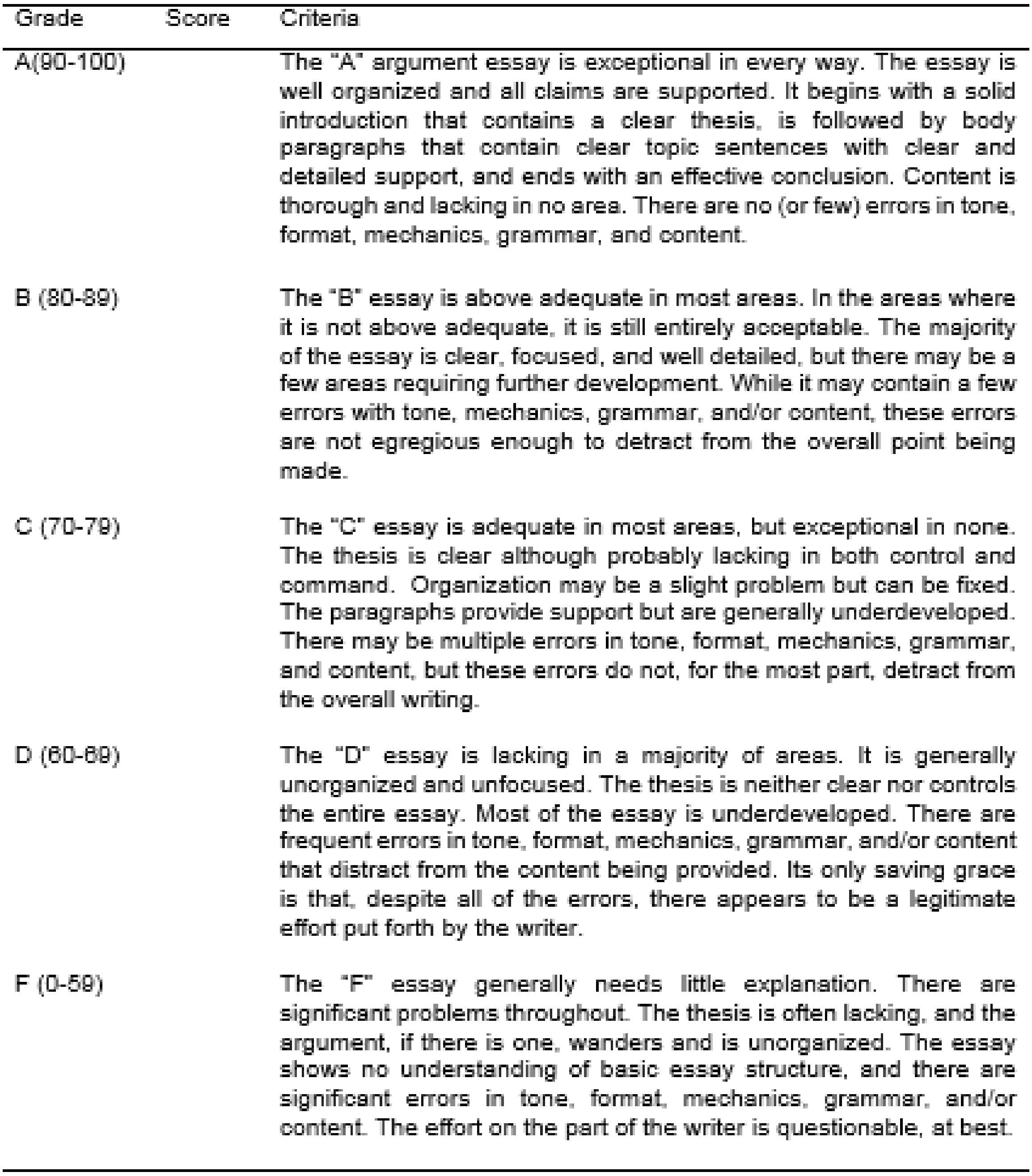
> *"Can write straightforward, detailed descriptions on a range of familiar subjects within his field of interest.*
> *Can write accounts of experiences, describing feelings and reactions in simple connected text.*
> *Can write a description of an event, a recent trip - real or imagined.*
> *Can narrate a story"* (Council of Europe, 2001, p.62)

As can be understood, both in terms of structures to be used and content to be included, to what extent B1 level students can write about the given topic is defined with CEFR descriptors and scales (Council of Europe, 2001). However, according to the type and purpose of the assessment, rubrics are commonly used instead of scales as they are more detailed. There are various types of rubrics like general and task specific rubrics, and holistic and analytic rubrics.

Holistic Rubric. Rubrics provide a guide to assess student performance and help raters to score marks in the examinations. The fairness of scores assigned to students is so crucial that in performance assessments giving a mark with a scoring rubric is out of question. Holistic rubric is a guide for assessment used to assess students' written process or production with pre-defined criteria including mechanics, grammar, organization, etc. (Gunning, 1998; Weigle, 2002). All the criteria related to one descriptive scale are expressed in the same paragraph. For example, there is only one band giving a general explanation about grammar and by which the rater judges both under achievers and very successful students (Moscal, 2000). Raters don't have to pay attention to each aspect of the written performance of students in detail and make an overall judgment of learners' writing ability as a proof of his linguistic ability (Wang, 2009). A sample holistic rubric (Aims Community College, 2018) is shown below:

Table 3

Holistic Rubric Sample by Aims Community College (2018)



Grade	Score	Criteria
A(90-100)		The "A" argument essay is exceptional in every way. The essay is well organized and all claims are supported. It begins with a solid introduction that contains a clear thesis, is followed by body paragraphs that contain clear topic sentences with clear and detailed support, and ends with an effective conclusion. Content is thorough and lacking in no area. There are no (or few) errors in tone, format, mechanics, grammar, and content.
B (80-89)		The "B" essay is above adequate in most areas. In the areas where it is not above adequate, it is still entirely acceptable. The majority of the essay is clear, focused, and well detailed, but there may be a few areas requiring further development. While it may contain a few errors with tone, mechanics, grammar, and/or content, these errors are not egregious enough to detract from the overall point being made.
C (70-79)		The "C" essay is adequate in most areas, but exceptional in none. The thesis is clear although probably lacking in both control and command. Organization may be a slight problem but can be fixed. The paragraphs provide support but are generally underdeveloped. There may be multiple errors in tone, format, mechanics, grammar, and content, but these errors do not, for the most part, detract from the overall writing.
D (60-69)		The "D" essay is lacking in a majority of areas. It is generally unorganized and unfocused. The thesis is neither clear nor controls the entire essay. Most of the essay is underdeveloped. There are frequent errors in tone, format, mechanics, grammar, and/or content that distract from the content being provided. Its only saving grace is that, despite all of the errors, there appears to be a legitimate effort put forth by the writer.
F (0-59)		The "F" essay generally needs little explanation. There are significant problems throughout. The thesis is often lacking, and the argument, if there is one, wanders and is unorganized. The essay shows no understanding of basic essay structure, and there are significant errors in tone, format, mechanics, grammar, and/or content. The effort on the part of the writer is questionable, at best.

Holistic scoring has been quickly accepted and widely used in assessing writing as it is considered applicable and flexible in assessments (Huot, 1990). The rubric choice is determined by the purpose of the assessment. When there needs to be a general understanding of students' progress and product, holistic rubric can be a good choice. As there is no need to make corrections on students' papers or writing some clarifying comments in holistic scoring, it is really favorable at the times of time constraint (Babin & Harrison, 1999; Wang, 2009). This type of rubric is really practical for experienced raters if they are familiar with the rubric, they can give a score in a short period of time (Wang, 2009; Yen, 2016). Being experienced as a rater may be quite advantageous, yet inexperienced raters can be good at scoring. In a study by Attali (2016), both newly trained and experienced raters showed similar results in the tests applied. The only variability in newly-trained group is in their scores for each band- higher or lower- although in their average scores, they don't show any significant difference.

As Jonsson and Svingby (2007) stated, 'holistic scoring is usually used for large-scale assessment because it is assumed to be easy, cheap and accurate' (p.132). In crowded classes or when the judgment needs to occur as soon as possible or only human raters are included into the scoring process even though the number of these raters is not enough, holistic scoring can be preferred. In the assessments where practice is more necessary than theory, holistic scoring is prominent. While reliability and validity of assessments are considered important, the practicality of assessment method cannot be denied. Indeed, most of the time an assessment's being practical is as essential as its being reliable. The advantages of holistic scoring are not limited to these. As Abeywickrama and Brown (2010) stated holistic scoring contributes to inter-rater reliability substantially and it makes it possible to use writing as an assessment tool in various disciplines. As holistic scoring is easy to apply and the descriptions used to score writing are clear enough to be understood by anyone, it can be preferred.

During holistic scoring, some underestimated factors can affect the scoring process and results accordingly. As Yen (2016) claimed, 'different raters may choose to focus on different aspects of the written product and they may be swayed by superficial factors such as length and appearance of an essay. And as it is possible for each writing product to appear just to a certain rater but not others, the examiner's mark may be a highly subjective one' (p.77). A huge number of researchers are on the same side in terms of holistic scoring's being subjective (Alanen et al, 2010; Hamp-Lyons, 1991; Kayapınar, 2014; Vaughan, 1991). In holistic scoring, while some raters focus more on to what extent the written performance is error free in terms of language use (Grobe, 1981), other raters pay more attention to correct use of vocabulary items (Engber, 1995) or how various vocabulary items are (Sakyi, 2001).

Raters’ assessment can be influenced by different aspects of writing, so this can result in variable scores assigned by different raters. In a study conducted by Wolfe et al (2016), a quarter of the variance in scoring resulted from raters’ consideration on how much students write and varieties in expressions. Considering all these, it is stated that holistic scoring is not appropriate to use when the decision hasn’t been taken on student writing yet and not enough to have an overall decision (Faigley, Cherry, Jolliffe, & Skinner, 1993). It is also claimed that holistic scoring cannot be used for all types of writing assessment as it doesn’t provide consistency within each category although the final mark is similar, which may prevent giving significant information about student performance. For classroom teaching and to support learning process, holistic scoring may not be adequate. Additionally, if raters aren’t trained well, holistic scoring cannot reveal accurate results (Abeywickrama & Brown, 2010). Raters not having a command of the items stated in each category in the rubric or not having enough practice with it may differ in scoring student writings. In a study conducted by Diederich, French and Carlton (1961), 300 essays were assessed by 53 raters, which shows a huge variety in scores and inconsistency among the raters as the reliability coefficient was a lot higher than the accepted. Such studies result in questioning the reliability of the assessment.

Huot (1990) believes that putting too much emphasis on reliability, but not taking validity into consideration is why holistic scoring is defenseless. The decision taken to assign a score is to be based on concrete reasons/criteria so that all raters assessing the same writing exams can give similar marks, which will increase reliability at the same time. Consequently, if there is more than one rater for each student performance to score using the same criteria, the reliability can increase. When each band illustrates specific features of writing performances, it is more possible to see how two different raters give similar scores by using holistic rubric (Siegert & Guo, 2009). Clear, understandable explanations are beneficial for each band showing the expected achievements of students. In the study by Wang (2009), eight raters scored students writing papers both holistically and analytically and the results showed that the raters gave consistent scores no matter which scoring rubric they used. That means it is unfair just to blame holistic scoring for the subjectivity of scores. On the other hand, for the type of examinations like proficiency exams according to which a fail or success of students has been decided, it needs to be questioned whether a holistic rubric should be preferred and students written performance should be assessed by an overall judgment regardless the number of students assessed. This issue requires to be undertaken in further studies.

Analytic Rubric. Analytic scoring is a way of rating by which every single aspect of students' performances is assessed in a detailed way. It is an assessment tool involving different categories for each aspect of students' performances, so there are sub-categories within each category (Yen, 2016). The categories and the appointed scores for each category can change depending on the exam types, learner needs, administrator purposes or curriculum (Brown, 2010). A sample analytic rubric (Rublee, M. R., n.d.) can be seen below:

Table 4

Analytic Rubric Sample by Rublee (n.d.)

Criteria	A	B	C	D/F
Organization	Clear organization that walks the reader through the paper, does not stray off topic	Clear organization but strays slightly	Organization is less than clear, or organization is clear but some digressions	Organization is unclear and/or paper strays substantially from topic
Argumentation	Paper has clear, strong arguments that go beyond description	Paper has discernable arguments but may be somewhat unclear or weak	Paper has arguments but often falls into description	Paper has little to no arguments, spends most time describing
Support	Numerous varied and relevant details and facts support arguments	Details and facts support arguments, but may not provide enough or may be as relevant as possible	Some details and facts to support arguments, but not enough and some lack relevancy	Little to no relevant details and facts to support arguments
Content Knowledge	Demonstrates excellent understanding of content and is comfortable with	Conveys content adequately but fails to elaborate	Gets basic content but is otherwise uncomfortable with material	Basic content is wrong, incorrect, or substantially incomplete

	nuances in material			
Originality	Demonstrates excellent analytical originality, either in creating new arguments or in relating facts in new ways(beyond what is covered in course material)	Demonstrates some, but not a great deal of, analytical originality, either in creating new arguments or in relating facts in new ways	Demonstrates little analytical originality, relies mainly on arguments and evidence already covered in class	Makes no attempt to provide original analysis
Level of Discourse	Variety of sentences, good use of cohesive devices	Some variety in sentence structure, adequate use of cohesive devices	Limited variety in sentence structure, little use of cohesive devices	Mostly single-clause sentences, little to no use of cohesive devices
Vocabulary	Precise diction, rich use of appropriate vocabulary	Generally good vocabulary choices with some variety, minor errors in diction	Limited vocabulary, not always precise or accurate	Incorrect use of vocabulary, very limited range
Grammar	No major errors, a few minor errors that do not distract	One major error or several minor errors that do not distract	Two or three major errors combined with minor errors	Numerous major errors

Assessing a second or foreign language writing process necessitates a valid and reliable judgment which was especially critical in the writing exams restricted with a duration limit (East, 2009). The scoring process of such exams should be organized carefully in order not to include the rating scale into the reasons like rater success causing unreliability problem. As raters infer students' ability in written language from what they write, rating process inevitably turns into a subjective evaluation. Therefore, it is necessary to make concrete judgments (Weigle, 2002) away from raters' individual decisions derived from their personal considerations. If they are not stated in the rubric provided with raters, students’ handwriting, weak use of spelling, the length of their written performance can be considered trivial details (Charney, 1984), which is mainly encountered in holistic scoring.

To increase reliability in scoring, raters may need a scoring rubric which doesn’t skip any criteria that can be considered important by raters. According to Jonnson and Svingby (2007), to increase the reliability in assigning a score to performance based assessment, using an analytic rubric with some sample scoring and trained enough raters is critical. In various studies, it is inferred that the discrepancy level among different raters is lower when an analytic rubric is used than a holistic rubric (Breland, 1983; Huot, 1990; Veal & Hudson, 1983). Much specifically, raters can agree with each other more easily while finalizing a score on a given student writing when they use an analytic rubric.

Writing is a multifaceted skill which not only includes students’ success in correct and variant use of grammar, vocabulary, spelling and punctuation, but also in organizing ideas in a coherent and meaningful way, in expressing the thoughts logically, etc. (Hamp-Lyons, 2005). For such an assessment, analytic rubric is essential as such a skill with lots of features to be regarded can only be scored using detailed criteria. In this way of assessment, raters need to share much time to evaluate each facet of written performance, which may make raters not to overlook any part. (East, 2009). While assessing writing, instructors may not focus on each student paper in the same way especially if they have to evaluate some of the papers at a different time or place. They might not feel or think in the same way, or they can pay attention to different qualities of the performance. Coherence and cohesion may overweigh grammar and accuracy or vice versa for some raters while some of them think content and organization are more important than those. Moreover, some raters can take legibility or sentence complexity into their consideration, whereas other raters assess writing just considering organization, content and grammar. When a rater assesses students’ writing in terms of content, not only the paper’s visual appearance involving criteria such as handwriting, layout, how long, how correct writing is written, but also how strong and diverse the words and language are chosen can be considered (Wolfe et al, 2016). If such criteria are not stated in the rubric provided clearly and in a detailed way, scoring difference comes out. In this sense, using an analytic rubric helps raters to stay focused and not to be lenient or severe in scoring.

Analytic rubrics hinder personal judgments so that assessment becomes more reliable. (Hamp-Lyons, 1991; Weigle, 2002). The point not to be missed out is that the criteria in the rubric can help as long as they are applied appropriately. (Janssen et al, 2015). When analytic rubrics are so detailed that any criterion needed is stated evidently on the rubric, it is expected that no variance in scoring will be noted, yet effective use of rubric is needed to hinder scoring difference. In a study conducted by Zhang (2016), raters were grouped according to their effective and accurate use of the rubric and it was found out that raters using the rubric precisely are good at merging what students write with the criteria, assess it accordingly and they are much more sure and confident compared to the raters who fail in assigning precise and careful scores.

Raters compare students’ performances with the previous ones and consequently they evaluate writing relatively. As each student writing doesn’t have the same qualifications, this kind of evaluation cannot be reliable without using a detailed, point by point scale. According to Goodwin (2016), raters give similar scores to the written performances of students when they are assigned by an analytical scoring as analytical rubrics are not prone to comments and interpretations a lot. In other words, raters’ inconsistency in scoring can be overcome using an analytical rubric.

Rater Consistency

Known as not only rater agreement, but also rater reliability, rater consistency is how consistent, correct and reliable scores raters assign to students (Yen, 2016). When raters use the same assessment criteria to score student writing persistently, reliable rating results are expected. Raters may not share the same opinions on the student performances. At that time, units responsible for testing need to find a way for an agreement on the score to be released (Johnson, Penny, & Gordon, 2000). How reliable a test measures can only be explained after it measures what needs to be measured. According to Huot (1990), there are four ways to demonstrate the reliability of an assessment: predictive, concurrent, content and construct validity. For performance assessments, predictive validity is important as the raters decide on a score by assessing the papers predicting how much successful they are. More clearly, raters have an impression on student success based on their performance and assign a score. This occurs especially during holistic scoring. Raters focus on some specific features of writing papers and determine a score in their mind. They may get impressed by ideas, content and organization, form and function, mechanics or the choice of words (Huot, 1990). Here it is necessary that each rater should get influenced from the same quality equally so that consistency among raters can be achieved.

Rater consistency is such a significant issue in especially performance assessments in which rater subjectivity shows up. While determining a score in a pass/ fail exam, giving an unreliable score can influence the students far beyond the expectations (Jonnson, Penny, & Gordon, 2001; Kayapınar, 2014). If the final score is inconsistent with student's real success after an assessment, it affects not only the students' educational and professional life, but also their psychology and faith in fairness.

When raters are not guided with clear objectives and criteria during an assessment, they would have instant decisions on students' performances (Lievens, 2001). These decisions are taken with inner considerations which aren't based on any criteria- holistic or analytic rubric. Raters do the scoring holistically without using a holistic rubric. The scoring process cannot be claimed to be reliable and valid unless raters are informed and provided with the same criteria to assess similar qualities according to the explanations stated on the guiding rubric at hand (Huot, 1990).

Among the various assessment tools like multiple choice tests, true-false, matching, fill-in; short answers, performance assessments like speaking or writing, no matter what kind of an assessment is applied, error is inevitable (Huot, 1990; Jonnson et al, 2001). The decision making process in assigning a score to an assessment, especially the assessments that are more prone to subjectivity, involves a high possibility of error. In scoring writing, independent rater or raters decide on a final score based on a scoring criteria. Raters assign a score using either holistic scoring or analytic scoring, which is decided by the testing committee of the department they work at or the management of the school. Depending on the rubric type, raters' scores can change especially for low level and high level student performances: raters' using analytical scoring give higher scores to lower proficiency writing performances, whereas when they use holistic scoring, the scores are high for higher proficiency writing performances (Zhang, Xiao & Luo, 2015). The final scores obtained from these ratings can be questioned by test takers or their parents if they get a score lower than they expect. What students ask for is a fair judgment during assessments (Wolcott & Legg, 1998). Scores shouldn't change depending on who assigns them, how and when raters assign them or which rubric is used. Errors may also result from students themselves. Students can be familiar with the topic given in the exam as they write a paragraph or an essay about it beforehand so that they can get a higher score than they can normally have in their own level; or if they have a terrible experience or get out of a routine just before the exam, their final score may not represent their real performance (Huot, 1990). Such errors deriving from students yield inconsistency. These types of errors which cannot be expected by testers are difficult to hinder; however, the errors stemming from raters need to be considered and so as to diminish the effect of error in assessment, reasons behind them are to be revealed in studies.

Most of the time it is difficult to differentiate the reason for the variety in scoring. That's to say, scoring differences may result from rater judgments or the mismatch of curriculum and student performance (Jonnson et al, 2001). The discrepant scores can emerge because of the internal inconsistency of the rater, external inconsistency of the rater or internal inconsistency of the student (Huot, 1990; Moscal & Leydens, 2000). Any of these reasons may jeopardize getting equitable scores assigned to students (Meier, 2012). Although some of the factors affecting both the scoring process and the scoring results can be overcome as they are predictable, some other factors, especially human factors need careful considerations.

Inter-rater Consistency. Reliability in assessment can be categorized into inter-rater reliability and intra-rater reliability. Inter-rater reliability is the extent of the agreement between or among multiple raters in scoring. (Siegert & Guo, 2009). While scoring student writing, different independent raters assign a score to the same student papers without knowing each other's score and without having an opinion about their judgments and the scores given by different raters are consistent with each other (Wang, 2009). Judgments and impressions of raters are parallel to each other when inter-rater consistency is ensured (Tashakkori & Teddlie, 1998). Including different raters who haven't been in such a study before in the scoring process makes both the result and the tools used in the study reliable as it reduces the prejudges of the raters, which helps to get objective conclusions (Marques & McCall, 2005).

The purpose of writing task can be understood differently by the teachers so there needs to be an agreement among them (Alanen et al, 2010). Rubrics are clearly helpful for raters to ensure inter-rater reliability (Jonsson & Svingby, 2007). Depending on the academic or managerial purposes of the school, a holistic rubric or an analytical rubric can be used by raters. No matter which rubric is preferred to assess student writings, so as to increase the inter-rater reliability, it is highly recommended that more than one rater should be included in the assessment process (Cherry & Meyer, 1993; Kane, Crooks, & Cohen, 1999). When just one rater judges the performance of a student, objectivity in scoring decreases because the extent of consistency in scoring when more than two raters are included in the rating process is believed to increase (Moscal & Leydens, 2000). As the main objective of using two or more raters in assessments is to get a result that best represents the students' real performance, the marking is anticipated to be more reliable. When multiple raters are assigned to assess student writing, they need to consider the same qualities of the performance as long as they can be consistent with each other (Huot, 1990). However, it is not always possible for the raters to correspond to each other while judging the same student performance. Wolfe at al (2016) states that the subjectivity of the scoring process is an undeniable fact causing variant marks for the same student performances scored by different raters. They may focus on different aspects of the paragraph or the essay, which can cause varying scores (Jonnson et al, 2001). This inconvenience in scoring may both trouble raters and harm the rating process. While assessing a student writing, raters can consider its content and organization, grammatical range and accuracy, mechanics, spelling and punctuation, lexical resource, coherence and cohesion and task fulfilment. Jonnson et al (2001) claimed that although the scoring result is consistent between or among different raters and it shows a good level of inter-rater reliability, raters' tendencies in focusing on a specific feature of student writing can also result in conflicting scores. While some raters pay a special attention to the content of student writing, other raters' primary goal may be to assess the same paper in terms of grammatical range and accuracy; or some raters are distracted with students' spelling and punctuation mistakes and tend to underestimate the other aspects of student writing, whereas other raters never mind such mistakes and just look for good grammar and vocabulary on the paper. Yen (2016) states that despite being asked to assess the same set of student writings with the same criteria, raters deliver varying scores as the way they perceive the qualities and judge them differ greatly. This difference can also be observed among experienced raters (McNamara, 1996). This situation means that variations among the scores are unavoidable (Kayapınar, 2014).

Rater effect is a never underestimated factor in scoring difference. Wolfe et al (2016) groups the possible reactions of raters as those giving consistently low or high scores, those having a tendency to give an average score to all papers, and those who couldn't assign invariable scores to the similar papers, which results in inconsistency. To abolish or at least to decrease the effect of subjectivity, summing or getting an average score of the ones assigned by different raters is commonly applied. Jonnson et al (2001) mentioned 5 different ways to arise the consistency of scores, which are: integrating different raters' scores, consulting an expert rater who is an instructor chosen beforehand and trusted when the raters cannot agree on a score and taking his or her final score into consideration instead of the mark given by the first assessors, blending the scores of both the first two raters and the expert rater, blending the score of the expert with the one that is the closest to the expert's, or having a discussion among the raters to come to an agreement on a final score. Among them, discussion takes a lot of time as the raters included into the discussion may not reach a consensus on the same points easily. According to Shavelson and Webb (1991), getting the average score of all the scores assigned by different raters has more reliable results than the score decided by just one rater. Sometimes, it is not even enough to take the average of the scores of raters. Calculating the raters' scores and the expert rater's score and getting their average is thought to be more reliable (Jonnson et al, 2001).

Another way to increase the consistency between or among raters is benchmarking. Instructors with different views come together to score some previously selected student writings and reach to an agreement on some standards before assessing real student papers on their own. The papers chosen to be scored need to be good representatives of the student writings to be assessed (Popp, Rayn, Thompson, & Behrens, 2003).

Marques and McCall (2005) gets the attention to the use of inter-rater reliability in studies and reminds that the amount of data to be assessed by raters shouldn't be more than they can handle as excessive amount of papers may lead to results not reflecting the reality. They also add that different raters need to be informed about the subject and according to what aspects they will assess student performances.

Intra-rater Consistency. Intra rater reliability is internal consistency of raters during their writing assessment. Raters experience inconsistencies that stem from being impressed by some aspects or factors that are internal to them instead of true aspects or different factors in students' written performances (Moscal & Leydens, 2000). According to Brown (2010), 'failure to achieve intra-rater reliability could stem from lack of adherence to scoring criteria, inexperience, inattention, or even preconceived biases' (p.28). Assessment process is a quite complex process in which predictable and unpredictable factors influence the final rating scores (Goodwin, 2016). Raters could be too harsh or lenient, their scores could be inaccurate or they might tend to give an average score or extreme scores especially when they compare the writing performances with the previous or following ones. They contrast students' similar writing performances, which might also change the scores given (Goodwin, 2016). Different from inter-rater reliability in which different raters assign similar or the same scores to the same writing performances and they come to an agreement easily, intra rater reliability means that once the rater is given similar writing performances and asked to score them, the scores the rater assign may change, which results variance in scoring (East, 2009). As a kind of open-ended response to the given topic, writing performances cannot be scored by more than one rater in many educational setting because of lack of raters or time, therefore intra-rater reliability is a big concern (Brown, 2010). Some external factors like raters' mood on the day when writing assessment is carried out, knowing which student's writing performance they are assessing and having biases to them, falling motivation or focus because of some constraints like time, place or the number of exams to be assessed influence the manner of raters and scoring results (Moscal &

Leydens, 2000). There could be many reasons for intra-rater inconsistency and Bachman (1990) stated 3 of them as making comparisons between the previous and following student performances, scoring sequence of writing exams, the criteria of the rubric given as they may be interpreted differently by raters causing ambiguity. It is asserted that to some extent rater scoring variance is acceptable as it is not possible to change raters' fixed personal traits (McNamara, 1996). As it is crucial to hinder any possible outside effect that can change the final score, it is suggested that assigning more than one score to the same performance or getting an average score of marks assigned in different time by the same or different rater can be effective. In a study conducted by Goodwin (2016), raters fail in understanding and interpreting the criteria stated on the given rubric in the same way. They may not make use of scoring rubrics enough as they have their own inner criteria and assign scores using these criteria. This results in halo effect (Park, 2008). One student writing performance may affect the rater and the rater can be harsher or more lenient to the next performance. Moreover, one aspect of a writing performance can be prominent in the assessment and that quality may influence the rater so that the rater can judge the other performances considering it. This happens mainly because of ineffective use of rubrics (Park, 2008).

Raters can be affected by the features of students writing performances, which results in prejudice or raised expectations on the following performances, so it might have negative or positive effects on the results (Hughes, Keeling & Tuck, 1983). Similar performances can get different scores because of their being judged differently as the focus of the rater for each performance can vary. Raters varying thinking process in similar writing performances results in intra rater variance, so in the study conducted by Zhang, (2016), using think-aloud technique, raters' assessing processes are analyzed and it is found out that the more accurate a rater's score is, the more they are aware of accuracy in their own rating process. As Brown (2010) states in his book, to increase intra-rater reliability, raters can question themselves thinking whether they are consistent in using criteria, they pay even attention to the criteria for each student performance, they don't include any personal criteria not stated in the rubric given, they apply the same considerations to all students' performances even they change their opinions in the middle of assessment, they read student performances more than once for consistency and to make sure, and they can stay away from any factor causing them to be tired during scoring.

No matter what rubric is used in scoring writing, assessment in writing tends to be subjective as it includes decision making process. The changes stem from not only students' performance itself, the scale used, different raters' different judgments, but also the factors resulted from the same rater. Some factors can be foreseen, whereas other factors are even far from prediction in the way that how much they manage to focus on the performance and how much they can merge the information with criteria (Zhang, 2016), their state of mind and physical wellness. All these aspects may cause the same rater to score students' writing performances which are mostly equal (Shohamy, Gordon & Kraemer, 1992).

Conclusion

Writing is such a skill that no matter how much or how detailed it is taught by instructors, its assessment is affected from various factors from students' own performance, types of writing exams to different raters' judging differences and raters' own judgment differences within themselves. As writing assessment is a performance assessment with no concrete answer, it is open to subjectivity and it is hard to achieve reliable scores especially when more than one rater joins the assessing process. Scholars' views on writing assessment methods, rubric use- holistic or analytical, rater consistency- among raters and within raters' own scoring- for the sake of consistent, reliable, valid scoring process and scores were explained in this section. In order to search for reasons behind scoring variance, how and where the study was conducted will be stated in the following section.

Chapter 3
Methodology

Introduction

So as to reach underlying reasons for scoring differences among raters and all internal or external factors affecting scoring results, 3 methods- questionnaire, think-aloud and interview- were used in this study and to demonstrate the way the study was conducted, firstly how the method was developed will be elucidated step by step through theoretical framework which explains three types of instruments mainly used in the study. Then, how the setting was decided and the participants were selected accordingly will be stated. Next the main features of the instruments will be explained in detail, and finally how the data was collected and analyzed will be declared.

Research Design

For this study, both qualitative and quantitative data collection methods were utilized because the aim of the study is not only to reach numerical data that later will be turned into statistics showing specific behaviors or opinions of participants, but also to try to deeper understand fundamental reasons and deductions for tendencies and behaviors. In other words, a mixed method research design was employed in the study because of three main reasons: to comprehend the aimed topic in depth, to confirm the results received from one instrument by comparing them with other results, and to ascertain the outcome of the study considering the assembled conclusions that were arrived using various methods, which is called ‘triangulation’ (Dörnyei, 2007). When the phenomenon is complex enough to be solved, using a mixed method can have many advantages as using more than one research instrument not only makes the study results more valid, but also helps the researcher to reach more reliable results by expanding the researcher’s point of view. What is more, triangulation is very useful for confirming the researcher’s judgments since it leads to look at the same topic from different viewpoints (Dörnyei, 2007). It cannot be assured that the results gained from different methods will match each other, however, it is expected that even these results will be clarifying, maybe leading to refreshing ideas.

As an initial instrument, a questionnaire was used to gain an insight of the participants' perceptions, level of knowledge and attitudes. By using a questionnaire, it will be possible to reach the number of participants planned to include in the study, which is almost impossible in other methods like interviewing especially when compared to this rate (Dörnyei, 2007). Furthermore, to obtain the information by this method takes less time than other methods, which is an undeniable fact. Notwithstanding, although it seems like an advantage for the researcher, it might turn out to be a drawback since the participants cannot contribute enough to the study in a limited time. They act voluntarily to participate in the questionnaire survey, but they may not focus on the questions, can hide some of their own ideas, or may get bored with writing full answers to open ended questions. To say, it has some poor sides that may generate lack of information.

For a complete understanding of the target issue, a second instrument, as an introspective method, think-aloud technique was used. The participants expressed their ideas verbally during the task and they were recorded with their consent. This technique helps the researcher to analyze 'the ongoing thoughts of the participants while they are focusing on a task'. (Dörnyei, 2007, p.148). What is expected in this process is to discover the inconsistency between the participants' thoughts and practices, if any. Here may show up another uncertainty: participants' focusing more on the think-aloud process than performing the task.

To be sure about their thoughts, the last research method, interview, was applied. With each participant, a face to face interview in single sessions was carried out. These were the semi-structured interviews using pre-determined questions and they were implemented just after the think-aloud sessions. Although the questions were prepared beforehand and the researcher guided the interviews, it was also hoped that the interviewees would expand the answers of questions or comment on some important issues so that fixed and superficial, not thorough answers can be avoided. The most important benefit of the interviews is that the researchers can gain an insight of different participants' considerations frankly and as a consequence, comparing each participant's opinions and judgments with each other. Using questionnaires, think-aloud technique, recordings and interviews construct a mixed method research - including both quantitative and qualitative research in both collecting and analyzing the data (Dörnyei, 2007). All these methods are really helpful in drawing a conclusion.

Settings and Participants

This current study was conducted in the School of Foreign Languages at a state university in Ankara, Turkey. This school welcomes more than 2 thousand students that will study in different departments each semester in the Preparatory Program. Students come from different cities of Turkey, even some of them are foreign students. Their ages change from 17 to 60 and more. In the departments of the students, the medium of instruction changes. The program for the students in whose departments the medium of instruction is fully (100%) or partly (30%) English is compulsory English. There is another group of students whose departments do not necessitate English as a medium of instruction, but they attend the preparatory program voluntarily to learn English. All students take classes for reading, writing, listening and speaking skills separately and they have a main course in which they mainly focus on grammar and vocabulary. At the end of the term or semester, if students whose levels are at least A2 level of English and above can collect 65 points out of 100 points during a term, they can take the proficiency exam. The proficiency exam has 5 parts: listening, reading, writing, speaking, and grammar. Students who get B1+ (high intermediate) are considered as having completed the English Preparatory Program. Proficiency writing section is evaluated out of 20 points, whose percentage contribution to the overall test is 20%. For this section, students are required to write an opinion paragraph on a given topic using at least 150 words in 45 minutes. Their responses in this task are assessed regarding four main criteria: task fulfilment, coherence and cohesion, lexical resource and grammatical range and accuracy using a holistic rubric. The rubric includes 5 parts: Good (20-17), Above average (16-13), Average (12-9), Below average (8-5), and Poor (4-1). Students are also penalized for irrelevant, off topic and not enough answer. Each part gives general explanations as a guidance for assessors. The assessors are the English instructors in the same university. They have different backgrounds, graduated from different universities; their ages range from 26 to 62 and their years of experience change from at least 3 to more than 25 years.

For this study, 12 students’ papers were chosen so as to be scored by 15 instructors decided beforehand. The papers of the students were taken from the English Proficiency Exam randomly. The main purpose is to eliminate or at least reduce the effect of any subjectivity as it is possible for raters to compare and group the writing exams as good, average, poor, etc. in their minds. Students were informed about the study and their consent was requested. 8 of the students were male; 4 of the students were female. These papers had already been scored and these students, who are the owners of the papers, were already placed to the classes suitable for their proficiency levels. The instructors as raters were tried to be chosen considering their years of experience. Among them there are 3 groups: first group included instructors whose years of experience are less than 10 years, second group included instructors whose years of experience are less than 20 years, and the last group included instructors whose years of experience are more than 20 years. Although there are more than 100 English instructors in different years of experience at the university, they were selected using Convenience Sampling procedure because this selection included classification, yet the most important criterion is the participants’ being convenient and volunteer to participate in the study. The instructors who were available and willing to participate were asked to contribute to the study and those who met the experience criterion were selected. Experience was not the only criteria. Their attitude and preferences or tendencies while grading student writings would also be analyzed. 3 of the instructors were male; 12 of them were female. This unbalance of gender is because of the limited number of male instructors at the institution.

Data Collection

Raters who would attend the think-aloud sessions to score given writing performances and also attend the interviews afterwards were selected among the participants attended the questionnaire through convenience sampling method. The 15 participants were divided into 3 groups according to their years of experience. Within each group, there were 5 instructors as raters having similar years of experience. Their ages also showed parallelism to their years of experience. They were unaware of the group they were included in. The writing papers to be scored and assessed by the instructors were selected randomly from those who were already assessed in the English proficiency exam. The instructors saw neither the names of the students on the papers nor the scores given to them beforehand by other instructors while grading the papers.

As a first step in data collection process, instructors were given a questionnaire survey prepared by the researcher. The questionnaires were distributed to 25 instructors by hand and they completed their surveys at the school. This helped to give clarifications on the points which were not clear enough for them. The goal of choosing a questionnaire as an initial instrument was to grasp a general knowledge about instructors' backgrounds, education life, teaching and assessing experience in writing, an understanding of their views on reasons for scoring differences and their attitudes towards professional development in their field. It would be a great advantage for the researcher to be aware of their participants' opinions on the related study before applying the main data collection method which is the think-aloud sessions with each participant. Another goal was to be able to compare what the instructors believe to know and practise and what they actually do during assessing students' writing exams. The instructors were given the questionnaires as a hard copy as this way was easier for them to handle. It took the participants 10 to 20 minutes to complete the survey. Almost in a month, the first data collection process, administering the questionnaire, was completed.

After implementing the questionnaires, the student writing exams that would be used to assess were chosen randomly and the students owning these exams were asked to give their consent by a written form for using their exams in the study. There were 12 student writing exams chosen. The reason why 12, not 10 writing exams were used was to diminish the possibility of not having enough papers to score. Some of the exam papers might be too short and thereby not contain enough information to score, which can prevent finding out the possible differences in scoring. It was also important not to lay a burden on the raters, so more than 12 papers might lead to lack of concentration and willingness.

Then it was time to decide on the instructors to include in the study as participants. In order to compare more results to draw a more reliable conclusion, 15 instructors were selected. As one of the research questions is about possible effects of the years of instructors' experience, it would be logical to split the participants into small groups considering this criterion. Selecting and grouping the instructors were brought about using Convenience Sampling. While grouping the instructors was categorized as they were planned to group according to their years of experience, selecting them among the whole instructors was carried out considering how suitable and willing they are and how much time they can share on this study. There were 3 groups, each having 5 participants having similar experience years.

Think-aloud sessions were conducted one to one. The raters were given a consent form to participate in the think-aloud session and an interview following it, 12 student writing exams to score and a holistic rubric they were already familiar with as they already used in proficiency exams. The instructors, as raters, were informed that they would be recorded while they were assessing the papers and this recording would be used in the study and later with their consent. Each think-aloud session took at least 20 minutes and the raters were not interrupted during the session. The more natural they did scoring; the better results would be reached. Some of the instructors asked to be alone during the recording, whereas the others let me be there and observe them. Some of the instructors requested to comment in their mother tongue as they believed they could express themselves better while some other raters asserted they could focus more on what they say if they think aloud in English. In order not to stay away from the purpose of the study and to get the best result, they chose the language themselves. There was a planned schedule showing on which day, at which hour each session would be conducted, so the instructors had already known when and what time they would join the study.

Just after each think-aloud session, as the last data collection instrument, interviews were carried out with the same raters. These interviews were also recorded within the consent of the participants. During these interviews, 5 questions were asked about the writing papers they had just scored. These questions were both general and specific questions, about the rubric, student mistakes, scoring style, any specific quality like grammar and accuracy, coherence and cohesion they especially pay attention to while grading the papers. The interviews were carried out just after the think-aloud sessions so that the raters didn't forget the experience and could answer any related questions with a fresh mind. Think-aloud sessions and interviews took almost two months to complete.

After collecting all these data, the results obtained from three different methods were compared and contrasted with each other and the prominent findings were analyzed in detail.

Instruments

In the current study, data was collected through three methods which are questionnaires, think-aloud sessions and interviews respectively. As the number of the participants for all data collection instruments was less than 30, it's become a non-parametric study. Firstly, to get the characteristics and personal opinions of the participants, a questionnaire which was prepared by the researcher by getting expert opinions was conducted. Later some of these participants were given students' writing exams to comment and score aloud and these processes were recorded by a camera. Finally, interviews were carried out with these raters just after the think-aloud sessions to understand their judgments and considerations on students' writings. If each research question is to be matched with a research instrument or instruments, it can be stated that for Research question 1, questionnaire results were analyzed with non-parametric Mann Whitney U test and Kruskal Wallis H and they were shown with tables. Some questionnaire results were also shared using bar and pie charts. Raters' opinions stated during the think-aloud sessions and interviews were gathered and used as quoted to compare the ideas; for Research question 2, questionnaire results were illustrated with charts, raters' judgments and opinions during scoring writing exams and during interviews were quoted and analyzed; for Research question 3, questionnaire results were analyzed, raters' judgments and opinions during scoring writing exams and during interviews were quoted and analyzed; for Research question 4, the scores assigned by the raters were analyzed and the mode, median and mean scores for each writing exam to be assessed were shared, the questionnaire results were shared by stating participants' ideas; and for Research question 5, the results of all research instruments were used.

Instrument 1. Questionnaire Administered to English Instructors. The participants' qualities, teaching experiences, whether they have any contributions to their own professional development, their habits and techniques while scoring student writing and their views on the reasons for inter-rater scoring differences were measured via a pencil-and-paper questionnaire. This questionnaire was developed by the researcher herself. The researcher studied on the issues to be asked, wrote the items and prepared a draft to be checked. Later by taking the opinions of two experts, the draft questionnaire was revised and made ready before conducting. This was done to ensure content validity. Any sentence or phrase that would be irritating to the participants, any statements that would lead to misunderstanding or absence of content were withdrawn. Moreover, some explanations were added to make items more clear. In this way, the questionnaire was finalized. Item wording is very important so as to get the best results. The questionnaire helped the researcher come up with various data. There were factual questions to learn about the participants age, gender, employment status, years of experience, level of education, level of English for listening, reading, writing, spoken interaction and production separately. Some behavioral questions were directed to learn their experience in teaching writing and the way they assess writing. Furthermore, participants were asked attitudinal questions to find out their opinions, preferences in scoring styles, attitudes towards writing evaluation and reasons for scoring differences.

The questionnaire included 35 questions which were both closed-ended and open-ended questions. In the first part of the questionnaire, there were questions involving multiple choice items and they asked academic backgrounds of the participants. In the second part, participants were asked about the time they spend teaching writing and the way they teach and practise using multiple choice items and yes-no items. Here, for clarification, it was given an empty space for additional opinions and comments. In the third part, yes-no items were used and participants were given an empty space if an explanation was necessary. Additionally, a question for numerical rating was preferred to see how they rank themselves in the given phenomenon. In the fourth and fifth part, open-ended questions were asked. These questions were mainly short answer questions while some of them were clarification questions and specific open questions. Open-ended questions provide the researcher with richer data and can make her consider the issue from different angles.

Instrument 2. Think, Comment and Score Aloud. As an introspective method, think-aloud technique was used as a second instrument in the study. 15 instructors carried out the tasks. They were selected considering their years of experience. They were asked to score 12 students' proficiency writing exams by using the rubric given. The rubric was the one they had already been using in the proficiency exams. While they were on task, they were recorded by a camera. The participants first were given detailed information about the conduct of the technique. It was necessary for them to focus on the assessing process, not on their wording or being recording. They were explained that they needed to keep talking while they were assessing and scoring the papers like talking to themselves and they were also requested to do the task as natural as they do in their professional life. Normally they needed to speculate in English, though some of the participants were reluctant to speak in English explaining that they could feel much more comfortable while speaking in Turkish, but cannot focus on the task as necessary while speaking in English. For this reason, they were free to use their mother tongue.

The reason why think-aloud technique was used for this study is to observe the raters while they are on task. They just revealed their thoughts and considerations at the same time. Raters' ideas were compared and categorized with thematic content analysis. The researcher can benefit from think-aloud technique since she has a chance to catch all details, even the ones the participants may forget when they are asked their comments after they score the papers. It is undeniable that some critics can be forgotten, however think-aloud technique gives the opportunity not to skip any critics of the raters.

Instrument 3. Interviews Administered to English Instructors. To understand the inner thoughts, judgments, reasoning and implications of participants, interviews were conducted as a last method. After raters finished scoring writing exams using the rubric given by thinking aloud, raters were asked questions about the scoring method they used during the assessment, their tendencies and focuses on students' writings. This was a semi-structured interview. 5 questions were addressed to the 15 participants who took part in the think-aloud sessions. The questions can be seen below:

1. Did you score students' writing holistically or analytically?
2. What is the reason for choosing the way you score?

3. Do you think the writing rubric given to you to score the writing exam is enough to guide you sufficiently, or have you had any difficulties in scoring because of the rubric?
4. What mistakes have you observed on students writing papers generally?
5. While determining the final score, what have you paid attention to most, in other words, how did you determine your final score?

As can be seen above, these were opinion questions about scoring style, what mistakes the raters observe on students' papers generally, and any specific criterion they especially pay attention to during scoring. Since the questions were open-ended, the participants commented on the questions easily, which could bring up new issues not estimated before. The questions were the guidance to direct the interview. The participants could mention any occasions they had in think-aloud session. All the comments and utterances of the participants were recorded so as not to skip any detail. Later they were transcribed so as to analyze. Raters' ideas were compared and categorized with thematic content analysis by focusing on the common themes in the utterances. From time to time, to encourage the participants to elaborate, the researcher used some techniques like staying silent to show more explanation is necessary, nodding to show that she is an active listener, listening to them attentively. To end the interview, the researcher thanked each participant for their participation. The table below illustrates which instruments were used for each research question:

Table 5

Data Analysis Summary

	Research Question	Instrument	Data Collection Sample	N	Data Analysis	Statistical Analysis
RQ1	Is there a significant difference in EFL instructors' scoring results in terms of their years of experience?	Questionnaire	English Instructors	25	Qualitative & Quantitative	SPSS Kruskal Wallis and Mann Whitney U tests
		Think-aloud		15		
		Interview		15	Qualitative	Descriptive
RQ2a	Is there a significant difference in EFL instructors' scoring results in terms of their use of rubric?	Questionnaire	English Instructors	25	Qualitative & Quantitative	Descriptive: percentage
		Think-aloud		15	Qualitative	Thematic Content Analysis
		Interview		15		
RQ2b	Is there a significant difference in EFL instructors' scoring results in terms of their familiarity with the rubric?	Questionnaire	English Instructors	25	Qualitative	Thematic Content Analysis
		Think-aloud		15		
		Interview		15		
RQ2c	Is there a significant difference in EFL instructors' scoring results in terms of their holistic and analytic scoring preference?	Questionnaire	English Instructors	25	Qualitative & Quantitative	Descriptive: percentage
		Think-aloud		15		
		Interview		15	Qualitative	Thematic Content Analysis

Data Analysis Summary (continued)

RQ3	Do the instructors have difficulty in scoring students' writing? If yes, what kinds of difficulties do they have?	Questionnaire	English Instructors	25	Qualitative & Quantitative	Descriptive: percentage
		Think-aloud		15		
		Interview		15	Qualitative	Thematic Content Analysis
RQ4	Is there a significant difference between/among the student grades when their writing exams are scored by more than one instructor?	Questionnaire		25	Qualitative	Thematic Content Analysis
		Think-aloud	English Instructors	15		
		Interview		15	Quantitative	Descriptive: mean, median, mode; percentage
RQ5	While grading students' writing, do the instructors pay their attention to a particular criterion more than others? If so, which criterion or criteria do they focus on more, and why?	Questionnaire		25	Quantitative	Mann Whitney U tests
		Think-aloud	English Instructors	15		Descriptive: percentage
		Interview		15	Qualitative	Thematic Content Analysis

Data Analysis

Although 3 different methods were used to collect data, the information attained via think-aloud sessions and the interviews following them was the core of the study. The research questions for this study were as follows:

1. Is there a significant difference in EFL instructors' scoring results in terms of their years of experience?
2. Is there a significant difference in EFL instructors' scoring results in terms of the rubric they use, their familiarity with and effective use of the rubric, their holistic and analytic scoring preference?
3. Do the instructors have difficulty in scoring students' writing? If yes, what kinds of difficulties do they have?
4. Is there a significant difference between/among the student grades when their writing exams are scored by more than one instructor?
5. While grading students' writing, do the instructors pay their attention on a particular criterion more than others? If so, which criterion or criteria do they focus on more, and why?

To find the answer to the first question, a questionnaire was firstly employed. In the questionnaire, instructors were asked for both their years of experience at the university where the study was conducted and the years of experience in total during their teaching career (Q.4, 5). They ticked the box illustrating their years of experience. A pie chart was used to show the ratio of the instructors for each experience group. They were also asked if the years of experience has a role in scoring differences in writing exams via questionnaire (Q.30, 31, 32). Participants' answers were compared and shared in quotations. The effect of raters' varying experience was also analyzed using SPSS, Kruskal Wallis Test. Not only their years of teaching experience, but also their ages, as a possible determinant of experience were included and analyzed with SPSS. Whether raters' fields of study were a factor in raters' assessment experience and their judgments were also searched and analyzed with SPSS Mann Whitney U test. Raters' educational backgrounds were also illustrated with a table.

Secondly, participants were chosen considering their years of experience and they were included in 3 different groups for the think-aloud sessions and interviews. After their remarks were recorded, they were transcribed and their similar and contrasting ideas were highlighted so that their considerations and scores were compared to see whether there is a significant scoring difference because of the raters' years of experience.

Whether instructors use a writing rubric or checklist was also asked in the questionnaire (Q.15, 18). As these questions were yes-no questions, the result was shared by giving the ratio of raters saying yes and no. The raters' familiarity with the rubric and their effective use of the rubric given were questioned both in the questionnaire (Q.25, 26, 27), and during the interview (Q.3). They were also given with percentages of the participants and the participants' comments were also shared with no addition in quotations. Raters' holistic and analytic scoring preferences were both asked in questionnaire (Q.15, 16, 17, 19, 20, 21, 22) and during the interview (Q.1, 2). Here, whether raters' scoring styles had any effect on the final score was investigated.

For the third question, possible difficulties raters might encounter during assessing and scoring writing exams were questioned with the questionnaire (Q.24). All difficulties that were experienced during scoring writing were also noted to see whether it has a role in different scoring results. Raters' comments and thoughts about difficulties they had in scoring were also given in quotations and they were compared.

As the same writing exams were scored by 15 different raters/instructors, whether different results between/among the student grades were attained or not was searched using the think-aloud technique. However, initially in the questionnaire, the raters were asked about the number of instructors who score the same writing exams for proficiency exams (Q.23) and the scoring differences among the raters (Q.28, 29). Their opinions were considered and if they had different points of view, these differences were used to illustrate what makes the raters score differently in the study. As to inter-rater reliability, it was observed that more data could be reached during instructors' commenting and scoring loudly. What instructors as raters did, told and speculated on the writing exams were recorded and transcribed verbatim so that the findings could be compared with one another. Similarities and differences were highlighted and later discussed.

While grading students' writing exams, it is quite difficult to stay objective, especially in determining the final score. To understand how consistent the raters can be in scoring, whether they have any priority or any quality they especially look for on students' writing exams like good command of grammar or content and organization, instructors were first asked in the questionnaire for their opinions (Q. 25). Later, how they score the exams were taken into consideration using the transcriptions of the recordings during think-aloud sessions. Finally, they shared their opinions on this issue when they were interviewed (Q.5).

Conclusion

In order to find out about whether there is unreliability in raters' own scoring and among different raters while scoring students' writing exams and if there is, how significant this unreliability is and the reason(s) for this, this study was conducted at the Department of Basic English of a state university with 15 instructors having different years of experience. Three methods were used: questionnaire survey, think-aloud sessions and interviews. At first, questionnaires were used to see educational, professional backgrounds and scoring styles of the instructors some of whom also participated to the sessions and interviews. Then instructors scored 12 students' proficiency writing exams using a rubric given and during scoring and assessing the papers, they were told to think and speculate loudly so that the camera recorded the whole sessions and their sentences were transcribed and compared with all participants' opinions. Finally, interviews were conducted with the raters to learn their considerations on the scoring rubric, student mistakes and their judgments. In methodology, how methods were developed, how the setting was decided, participants were selected, data was collected and analyzed were stated in detail. The findings acquired from these methods were analyzed and the results were stated in the subsequent sections.

Chapter 4
Findings

Introduction

After a detailed data collection process, the results gathered on scoring differences among raters when written performances are assessed by more than one instructor will be stated. In this section, the results of the study and the findings of each research question are shared and analyzed. In this study, three different methods were applied: questionnaires, think aloud sessions and interviews. The number of participants for the questionnaire is 25 which involves 21 female 4 male participants. Their ages change between 20 and 65. 16% of the participants are older than 51, 52% of the participants are between 41 and 50 years old, 24% of the participants are between 31 and 40 years old and 8% of the participants are between 20 and 30 years old. When the participants are compared in terms of age, it's easy to say that most of the participants are older than 40 years old. For the second method -think aloud sessions-, 15 participants were selected. 12 of them were female and 3 of them were male. They were divided into 3 groups in terms of their years of experience as Group 1 with more experienced raters, Group 2 with experienced raters and Group 3 with less experienced raters. The same 15 participants also attended the interviews.

Results of the Research Questions

Research question 1: Is there significant difference in EFL instructors' scoring results in terms of their years of experience? So as to find an answer to the first question, the participants were initially asked to share their years of experience in teaching in general and at the university using the questionnaire. The results can be seen in the pie charts below:

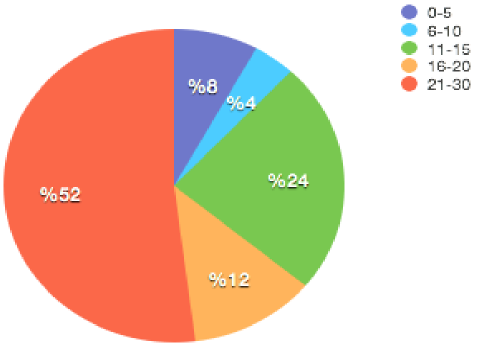


Figure 1. Years of experience in teaching

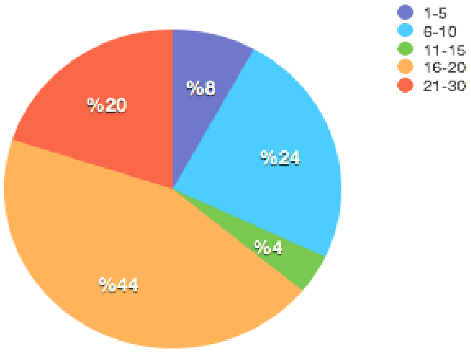


Figure 2. Years of experience at the university

As seen in the charts, more than half of the participants' years of experience in teaching is between 21-30, whereas the highest ratio in the participants' years of experience at the university belongs to those with 16-20 years of experience. So as to see whether raters' teaching experience has an effect on rater inconsistency, via SPSS, a Kruskal Wallis test was used. The table below shows the result of the test:

Table 6

Rater Experience

Experience	N	Mean Rank	X^2	p	Median
High	5	6.70	7.6139	.280	2.0000
Medium	5	6.70			
Low	5	10.60			
Total	15				

A Kruskall Wallis test was applied to see any statistically significant difference among the raters grouped in terms of their years of experience. The result revealed insignificant group differences in experience attributions: for high group N=5, medium group N=5, and low group N= 5. Total mean of the scores in experience attributions X^2 (2, N15) =7.6139 and the Asymp. significance, p=.280., which is not statistically significant. The median score of the scores is 2.000.

Raters' ages were also taken into consideration as there may be a connection between age and experience in teaching and assessing and to understand whether age is a factor in scoring difference, A Kruskall Wallis test was used and the result can be seen in Table 7:

Table 7

Rater Age

Age	N	Mean Rank	X^2	p	Median
41-50	5	9.50	7.6139	.678	2.0000
31-40	7	8.29			
20-30	3	6.70			
Total	15				

A Kruskall Wallis test was revealed insignificant group differences in age attributions for all groups: high group N=5, medium group N=7, and low group N= 3. Total mean of the scores in age attributions X^2 (2, N15) =7.6139 and the asymp. significance, p=.678., which is not statistically significant. The median score of the scores is 2.000.

As for their educational background, among 25 participants, 12 of the participants have a bachelor's (B.A.) degree in ELT, 7 of them have B.A. in English Literature, 2 of them have B.A. in English Translation and Interpretation, 1 of them has B.A. in English Language and Culture and 1 of them has B.A. in American Culture. 1 of the participants has a doctoral (P.H.D) degree in Curriculum and Instruction, 8 of the participants have a master's (M.A.) degree and half of whom have M.A. in ELT while 1 of whom has M.S. in English Literature, 1 of whom has M.A. in English Language and Culture and 2 of whom have M.A. in other departments. 8 of the participants who had B.A. as the highest qualification in departments apart from ELT had a teaching certificate in English Language Teaching after they graduated. All this information is visualized in the table below. ELT stands for English Language Teaching, Literature stands for English Language and Literature, Linguistics stands for English Linguistics, Translation stands for

English Translation and Interpretation, English Lang. & Culture stands for English Language and Culture, American Culture stands for American Culture and Literature:

Table 8

Educational Information of the Participants of the Questionnaire

Fields	B.A.	M.A.	P.H.D.	T. CER.
ELT	12	4		8
Literature	7	1		
Linguistics	2			
Translation	2			
English Lang.& Culture	1	1		
American Culture	1			
Other		2	1	

The role of instructors' changing fields and their possible effects on any significant difference in scoring difference was looked for via SPSS with the help of a Mann Whitney U Test, the result was shown in the table below:

Table 9

Raters' Field of Study

Field	N	Mean Rank	Sum of Ranks	U	Median	Z	P
Elt	9	8.11	73.00	26000	1	-.118	.906
Other	6	7.83	47.00				
Total	15						

As the number of the participants is less than 30, for this non-parametric study, a Mann Whitney U Test was conducted to show the effect of the fields of participants on any scoring difference. The test revealed an insignificant difference in field attributions of ELT graduates (*Median* = 1, *n* = 9) and the graduates of other departments (*Median* = 1, *n* = 6), *U* = 26000, *z* = -.118, *p* =.906, which is less than .01.

Instructors' experience in teaching writing was also searched for through the questionnaire. In Part II, the participants were asked how many hours they teach English and writing in English, how they teach, whether they teach skills separately, whether they give a writing checklist to the students, and whether they give feedback to the students during or after writing. 76% of the participants stated that they teach English 19-25 hours a week, 8% of them said more than 25 hours, 4% of them said 12 hours, 8% of them said 13-18 hours. Among these hours, 72% of the participants said they spend 1-3 hours teaching writing in a week and 20% of them spend 4-5 hours and 8% of them never teach writing in a week.

76% of the instructors who participated in the questionnaire stated that they teach skills in an integrated way while 12% of them said they teach skills separately. They were also asked how they teach and the answers varied:

- teaching organization and language use and going over sample paragraphs
- after teaching necessary skills, making them write 1st , 2nd , final drafts
- by following the book and materials
- teaching vocabulary and grammar first, organization later
- firstly, by semi-guiding and let them write by themselves
- through sample sentences, paragraphs and practicing together
- theory-controlled practice- produce
- by pre, during and post writing

Generally, participants showed a tendency to follow the materials and the course book supplied by the curriculum office. Some of them stated their teaching depends on the level of the students, but they usually do practice together with students. In terms of giving feedback to students, all the participants stated that they give feedback to students orally or in written form, sometimes during, but generally after writing. Some of them make corrections on the paper while there are participants who clarified that they avoided marking the mistakes on paper. Instead, they just write a short comment and students can ask questions if necessary.

With the questions 30, 31 and 32, participants were asked directly about the role of experience in scoring difference, their first year experience in scoring students' writing and whether they had any changes in terms of scoring as they got experienced more. For question 30 'I think the year of experience an instructor has a role in scoring differences in writing', 68% of the participants agreed that years of experience is one of the reasons in scoring while 32% of the participants said no.

So as to understand the role of experience, participants were directed the question of what mistakes they would make while scoring in their first year of experience. What they mentioned on this issue is:

- Giving much higher or lower marks
- Focusing on grammar, ignoring content
- Skipping some mistakes because of not understanding the criteria and not applying it.
- Being impressed by students' grammatical structures and effective use of vocabulary
- Being very strict while marking
- Tending to be more analytical, so focusing on the mistakes rather than the strengths

Additionally, the participants answered the question whether they have changed the way they score students' writing as they get more experienced. 8% of the participants said they didn't change the way they score, 16% of them didn't comment on this question and 76% of the participants noticed they and their marking changed in time. In what ways they changed the way they score were stated in the questionnaire as follows:

- Being dependent to rubric
- Using the holistic scoring
- Paying attention to all aspects, not just one
- Evaluating the papers in a more analytic way
- Grading the papers higher if they communicate well

- Tolerating minor mistakes
- Scoring in a shorter time
- Focusing on the message

Experience contributed to instructors in different ways. Some of them stated that they score the papers more holistically while some of them do it more analytically. Some raters claimed that they are inclined to use rubrics more. The most common change was observed in focus on grammar. Raters mostly concentrate on content as they get experienced. One rater also stated his/her opinion on the role of experience during the interview:

> '*When I started teaching almost 26 years ago, I was too strict while marking papers. But now I think I am more flexible because grammar mistakes, I mean we all make grammar mistakes but if you communicate the message, it doesn't matter then whether you have minor grammar problems or vocabulary problems.' (T15-interview)*

Research question 2: Is there a significant difference in EFL instructors' scoring results in terms of

a) their use of rubric. Initially, the role of rubric use in scoring difference and the effect of using or not using a rubric while especially scoring critical pass or fail exams were investigated via the questionnaire. In Part III, 3 questions were directed to the participants. They were requested to put a tick to the suitable box. For the question 'I think using a rubric is crucial in scoring writing', 92% of the participants said yes, agreeing on the importance of rubric use and only 8% of the participants said no clarifying that rubric use depends on the level of students. Some of the participants who said yes added that using a rubric helps them to get a general understanding as a general guide so it doesn't need to be detailed and content should be given more priority in the rubric. To what extent they think instructors should abide by the rubric during scoring was the next question concerning their rubric use. 4% of the participants said they didn't think every instructor should abide by the rubric provided, 12% of the participants partly agreed claiming that it depends on the experience of the scorer. If a scorer is experienced, s/he doesn't have to abide by the rubric. 84% of the participants stated that every instructor should obey the rule of using the rubric given.

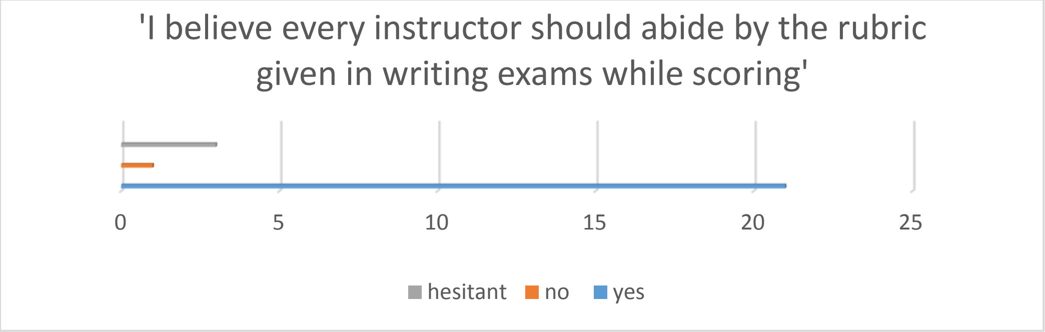


Figure 3. Item 18 of the questionnaire

As a last question, the participants were asked to consider themselves out of 10 in terms of how effective they use the writing rubric given to them. The table below shows raters' self-considerations. Out of 10 points, no rater gave a score between 1 and 4 points to themselves. 4% of the raters scored 5, 4% of the raters scored 6, 12% of the raters scored 7, 48% of the raters scored 8, 16% of the raters scored 9 and 16% of the raters scored 10 out of 10 points in terms of their effective rubric use. It was clearly observed that most of the raters considered themselves almost effective as they gave 8 out of 10 points.

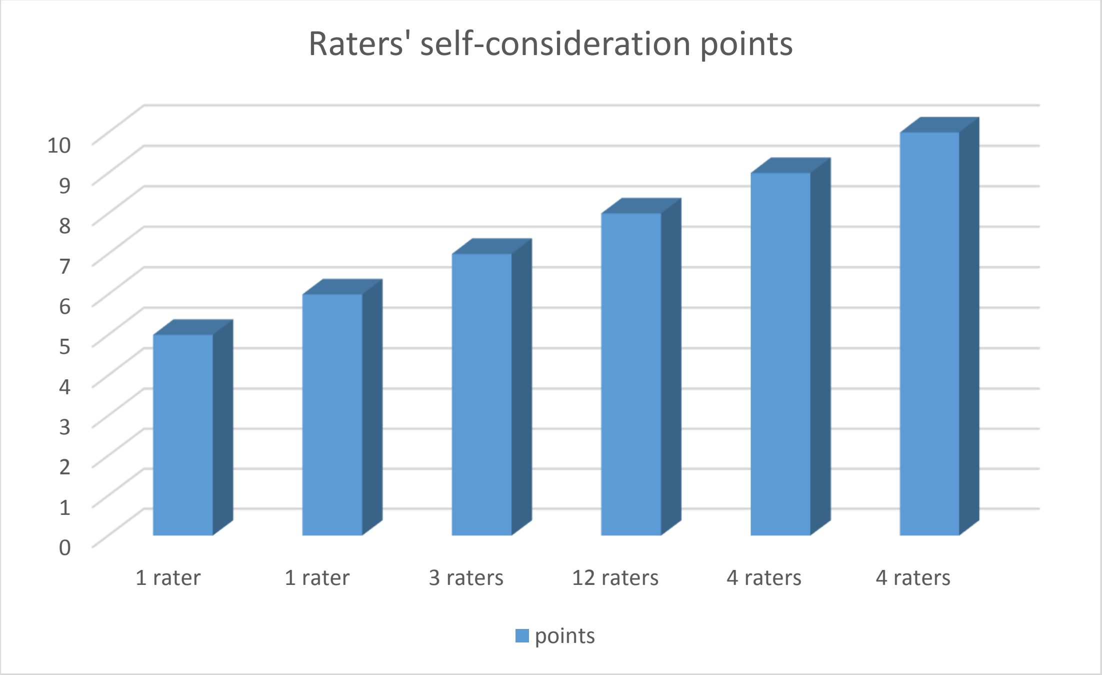


Figure 4. Item 27 of the questionnaire

During scoring the papers, it was observed that some raters sticked to the rubric given to them, whereas other raters ignored it and gave their scores in their minds. The rubric provided to the raters was holistic including 5 bands: good, above average, average, below average and poor. Each band involves clarifications on language and vocabulary use; tone and mechanics; organization and content. Apart from these criteria, there is an ‘additional considerations’ section which leads the raters to how they should score the papers in different situations like having no response, totally irrelevant response, personal opinion not stated, or multiple paragraph/ essay format. Maximum grades for such situations are provided so that there won’t be a huge gap between the grades. One question that needs to be considered is how aware raters are of this section. While some raters know that they need to take off 8 points if a student writes multiple paragraphs instead of a single paragraph and apply this rule, some other raters don’t prefer to cut off any points as they think 8 points is too much to take off in proficiency exams and they refuse to do it or they just cut off 1 or 2 points. Actually, even this number changes from rater to rater as some can cut off 3 or 4 points, but others can prefer 1 or 2 points. On the other hand, there is another group of raters who know the wrong information about how many points they need to cut off. Instead of deciding on a score out of 12, they wrongly assess those papers out of 16:

> *‘… the student needed to write one paragraph, but he or she wrote 3 separate paragraphs. But as this is a proficiency exam, I’m against taking off points for this reason. But of course some points should be taken off.’ (T8)*
>
> *‘…it has multiple paragraphs, not a single paragraph. For this, maximum score is given as 12 in the rubric.’ (T6)*
>
> *‘…he wrote multiple paragraphs, so I will assess it out of 16.’ (T1)*
>
> *‘the student didn’t write a single paragraph…so I can score it as 10 out of 20’ (T3)*
>
> *‘I normally don’t care about whether the student wrote a paragraph or an essay as it is a proficiency exam. Students may ignore it, forget it, or are not careful about it.’ (T9)*

> *‘it is written in an essay format, but I keep in mind that the exam was the proficiency exam and the students may have not leant how to write before, but they understood the topic and supported it. So it can be 15 or 14, but 15.’ (T11)*

> *‘the student wrote in an essay format, but I don’t care this a lot. Actually by thinking that all raters cut off some points for this, I can also cut off 1 or 2 points.’ (T9)*

As understood from the statements above, there are different ideas in how to consider the papers written in an essay format. Some raters assess multiple paragraph format papers out of 20 while others assess them out of 16 and some just score them as stated in the rubric knowingly and unknowingly. The points cut off for this rule may differ from 1 point to 8 points, which creates a 7-point- difference. Some raters even never cut off any points by thinking that it was the proficiency exam. Below there was an example of how a rater assesses a writing exam not written in a paragraph format:

> *‘My score for this is 12. If the student had written in a paragraph format, the score would have been higher.’ (T8)*

The rater is required to score such a paper out of 12, normally; however, as s/he also stated ‘the student didn’t use various vocabulary items, he or she had grammar and spelling mistakes and there were some incomplete sentences’, the score was 12 out of 12. Another striking point is about scoring the papers which are needed to be assessed according to ‘additional considerations’ rules.

> *‘…this paper is totally out of topic. So when we consider this out of 4, the structures he or she used are not bad actually, so my score is 3.’(T3)*

> *‘…this paper is totally irrelevant response, so it is 4.’ (T5)*

> *‘The student totally misunderstood the topic…the paragraph is irrelevant. The student just mentioned and praised his or her family, so it could be 6 or 7.*

The student never mentioned staying alone or with friends. My score is 8.' (T8)

'This paragraph is out of topic and it is a multiple paragraph. The student has her point, but there are problems in grammar, punctuation, so it is going to be 1.'(T13)

'This is out of topic, the student didn't answer to the given topic, so my score is 1 for this paper… not only off topic but also too weak in stating the opinions.' (T9)

For the same student paper which has totally irrelevant response and needs to be evaluated out of 4, raters have different ideas and implementations of the rule in scoring. The same student can get 1 or 8 for the same performance. While some raters just give 4 out of 4 as they suppose that when the answer is irrelevant, without assessing the paper in terms of organization or language use, they should assign a score as just 4, other raters may be aware of how they should assess those irrelevant papers, but there are also raters who can give a score more than 4 considering language use or organization of the student and never mind the rule for off-topic-writers. As understood, there are multiple implementations of the rubric and multiple assessment types by raters with different years of experience.

'…so we should fully focus on the criteria provided with.' (T13)
'I usually assess the papers overall, holistically, but when I hesitate to put a paper into the suitable band like poor, average, below average, etc., I always use the checklist in front of me.' (T9)

Raters can also have different approaches in using the writing rubric given to them. A rater from Group 1 (most experienced) tended not to use the rubric unless it was difficult for them to give a final score easily and although the rubric is holistic, raters assert they score holistically without a rubric, even a holistic one, whereas A rater from Group 3 (less experienced) looked at the checklist each time to make sure. G3 raters were also careful about other 'additional considerations.

> *'…overall it is going to be 10 out of 12. If it were a kind of paragraph, she would get a better score than 10, but it has to be evaluated out of 12.' (T13)*

As seen above, the rater scored the student paper out of 12, not 20 because the student didn't write a single paragraph, so some raters pay attention to what is written on the rubric, while other raters may not take it as serious as they do.

b) their familiarity with the rubric. In this study, there are various instructors with different educational backgrounds, years of experience, experience of different institutions and their applications. Not only raters who started to work in the school where the research was made or who just started working as a brand-new instructor, but also raters who were about to retire joined the study. Therefore, raters' familiarity with the rubric may change depending on the situation. Below there is an example of a rater who is among the less experienced raters (Group 3):

> *'I will evaluate the writing criteria because I am not familiar with this. First I will read the rubric and then score the papers… A few errors? How do I decide on a few errors?' (T5)*

The rater didn't have much experience in teaching and scoring writing and just started to work as an instructor, therefore s/he was not familiar with the rubric. However, another example could be given to show what the reaction of a rater who was in experienced raters (Group 2) was when s/he used the rubric for scoring:

> *'I'll first read the rubric. I'm familiar with it.' (T1)*

Some raters can't get used to the assessment process easily and they need to score some papers so that they could score in a more relaxed way. Before they feel ready to score the papers, they needed to assess the student papers twice or three times to make sure. T11, T10, T9 are among these raters. When they score the papers immediately, without getting ready and prepared to start scoring, they may encounter some problems and have some inner inconsistency. Some G2 and G3 raters talked about their familiarity with the rubric during the interview as follows:

> *'I am used to assessing students separately, according to each category. Later I collect all the scores.' (T 12-interview)*

‘I needed some separate items for structure, grammar, maybe content.’ (T5-interview)

Here the raters mean they assess and give a score for organization, content, language use, mechanics of the papers separately and count them and finalize a score out of 20 as they are not familiar with the rubric and found it difficult to score the papers with the rubric provided. These raters couldn’t give up the use of the previous rubric which was applied in the school and which they were really used to.

Another statement which belongs to a rater among more experienced ones (Group 1) also showed how familiar the rater is:

‘’Controlling idea or ideas given in the prompt not mentioned?’ I couldn’t get a clue what they mean.’ (T9-interview)

Considering the statement above, it can be said from all groups, there are raters who are unfamiliar with some criteria stated on the rubric. The unfamiliarity sometimes shows up when raters experience inconsistency. For two different papers, the same score was assigned by the same rater. Here are the comments of the rater for these two papers:

Paper 5: ‘The student knows how to write a paragraph; he couldn’t write well. He had basic grammar and vocabulary errors. He couldn’t express himself enough, so it is 7’

Paper 6: ‘The student couldn’t produce enough ideas. The biggest problem is in content. There aren’t many mistakes hindering the meaning. Compared to paragraph 5, its level of English is better, so let me give 7. ‘(T11)

For paper 5, the rater thinks the student has problems in language use while for paper 6, s/he says the problem is in content, not in language use and s/he finds paper 6 better than paper 5, but both papers were scored as 7. The same score was assigned for one paper just because of grammar. And for another paper, the same instructor assigned the same score because of lack of content:

‘…it is a free writing, so I will assess it in below average band.’ (T6)

> *'This paper is also like free writing, unorganized…the student used the language well... the student doesn't know the basic paragraph organization, so it is 4.' (T6)*

Below average is a score between 8 and 5 points according to the rubric. This rater decided to give a score in the below average band and the poor band because of the same reason: lack of organization. The focus here is organization, not grammar or content. Normally, in the rubric, there was no section explaining how many points needed to cut off when the student lacks of organization or content, the rater decided to assess the paper in the below average band or a lower band, which is his/her own decision. Actually, in the previous rubric which had been used by the raters for years, there were 3 sections: content and organization (10 pts.), language use and mechanics (6 pts.), vocabulary (4 pts.). As some raters were used to it, they tended to cut off a lot of points for lack of organization and they used this rubric instead of the rubric given to them unconsciously or consciously as they thought that rubric was quite well and they used it many times.

> *'the student wrote in a free style. Additionally, there are lots of spelling and grammar mistakes, so my score is 9.' (T4)*

For a paper written in a free style, another rater's score was 9, as here the focus of the rater was not only on organization. Again, since there was not a specific point to cut off for a free writing paper, raters decided on a point in their mind and applied it.

Although some criteria are stated on the rubric, raters are just unaware of them or some criteria are not stated on the rubric, they suppose they are there and they pay attention to them while scoring the papers.

> *'Okay there are spelling errors, but I think there is no need to take off any points because of spelling errors.' (T6)*

In the rubric the raters were provided with, tone and mechanics (spelling and punctuation) were among the criteria although some raters take them into consideration, others don’t. They can’t be sure whether there is a criterion about spelling errors or not in the rubric.

In the questionnaire, to see the importance of familiarity with the rubric in the eyes of raters, they were asked whether using an already prepared rubric is more effective than developing their own rubric. 36% of the raters stated that developing their own rubric is more effective, but 64% of the raters believed using an already prepared rubric is more effective. Some raters added that using an already prepared rubric is important for the sake of students’ equal chance in scoring.

Raters’ familiarity with the rubric and if they had any difficulties in scoring writing exams because of the rubric were also questioned in the interviews. What they stated can be seen below. The opinions of raters having difficulty in scoring are under ‘Negative’ column and raters’ opinions on effective rubric use are shared under ‘Positive’ column. The opinions of raters who were hesitant about the use of the rubric given were shown under ‘Neutral ‘column.

Table 10

Instructors’ Opinions on the Writing Rubric

Positive	Negative	Neutral
I like the way it just divided the scores, I feel safer.(T15)	I have difficulty in evaluating holistically because when I look at the paper, I’m searching for some categories for grammar, organization, etc. (T12)	Neither good, nor bad. There is no perfect rubric. (T14)
The rubric leads us throughout the paragraph. (T13)	I’m not happy with additional considerations and never applied any of these. Personal opinion not stated? What does that mean? (T9)	I don’t follow it. Experience says more than the rubric. (T7)

Too detailed rubrics are not any use for me. If you decide everything on the rubric, it will mislead you. (T11)	Not enough because it says 'the text sufficiently addresses the prompt' how do we define sufficiently? It says 'a few grammar errors' how many errors? 1, 2? We need a more detailed rubric. (T5)	It's OK in general, but for a multiple paragraph with a good language use and vocabulary and good supporting sentences, I had difficulty. (T6)
Clear criteria, clear additional considerations (T10)	For proficiency exam, I would like a more detailed rubric.(T2)	I have some rubric in my mind. Grammar and organization are not my main criteria.(T4)
It gives everything in detail (T8)	If I had an analytic rubric, I would be faster. With holistic rubrics, I just go back to papers and read again and again and compare them, but it is OK. (T1)	Good, but after deciding on a band, I can't give a score easily. I can't decide on to give 12 or 9 in the average band, so I compare the papers. (T3)

Considering the table above, the ratio of the raters in positive, negative and neutral side are the same. Raters not having difficulty in scoring because of the rubric stated that the criteria, definitions and the divisions of the scores help them to assess in a safe way and it guides them enough to assign a score. 2 of the raters were experienced raters (Group 1), two of them were less experienced raters (Group 3) and one of them is in experienced raters (Group 2). However, raters who were not satisfied with the rubric expressed several reasons for having difficulty in using the rubric provided. Not having a specific category showing points for a specific criterion is one of the reasons. To illustrate, to assess a paper in terms of organization, some raters want to know that they need to assess the paper out of 10 points, or they need to assess content out of 5 points. Some raters may have difficulty in understanding the explanations for scoring bands on the rubric. T5 claimed that clarifications to guide the raters were not clear enough. Below the table shows an example of a band showing above average performance:

Table 11

A Scoring Band on the Rubric Raters Use

Above Average: 13-16
The paragraph is above adequate in most areas and exceptional in some. In the areas where it is not above adequate, it is still entirely acceptable. The text sufficiently addresses the prompt. The majority of the paragraph is clear, focused and well-detailed, but there may be a few areas requiring further development. While it may contain a few errors with grammar, use of vocabulary, tone and mechanics (spelling and punctuation), these errors are not drastic enough to detract from the overall point being made.

In the table 11, the above average band is shown with its descriptions. According to the rubric provided with the raters, a writing exam that is considered in this band needs to fulfill the requirements of a paragraph in most areas and be notable in some areas like vocabulary or organization. Even if there is something lacking, it is still accepted in this level. What is written as a response to the given topic is relevant and sufficient enough with details and clear clarifications despite some statements that need to be explained more. Errors in grammar, mechanics, and vocabulary use can be acceptable as long as they don’t hinder the intended meaning.

Some adverbs like sufficiently and adjectives like a few, adequate were found unobvious for some raters, so they have difficulty in having a decision. They find the explanations given to clarify these statements not enough to guide them. The same situation was recognized for another rater (T9). S/he expressed the explanations were ambiguous in ‘additional considerations’, so s/he didn’t apply any of them while scoring.

Table 12

Additional Considerations on the Rubric Raters Use

	Maximum Grade
No response	0
Totally irrelevant response	4
Controlling idea/s given in the prompt not mentioned	12
Personal opinion not stated	16
Multiple paragraph/ essay format	12

The table above includes 5 situations in which what maximum score raters need to assign is shown. If a student doesn't give any response to the given topic, the score should be given as 0. If the answer of a student is irrelevant to the given topic and the student just mentions other issues apart from the given topic, the paper should be assessed out of 4. If a student writes about the topic, but changes the controlling idea and writes different statements that are not related to the main idea expected, his or her writing exam should be assessed out of 12. If a student doesn't share his or her own opinions, examples, explanations with enough support and uses memorized phrases, his or her exam will be evaluated out of 16 and if a student doesn't write a single paragraph, but writes multiple paragraphs like an essay, the exam will be assessed out of 12.

It was asserted that some expressions like 'Personal opinion not stated' were unclear by justifying that what a student answered in the exam paper was the student's own opinion. What is more, a rater had difficulty in scoring a paper written in an essay format revealing that a paper with a good language use and content was to be evaluated out of 12, which means 8 points were cut off just because the student wrote separate paragraphs. This situation made some raters hesitant whether the rubric was good or not as T6 stated. Another reason reported by the raters was the need for a much detailed rubric especially for proficiency exams. Raters claimed that holistic rubric is too general to help them direct to a final judgment; moreover, it takes more time than necessary since raters need to reread the papers and rubric to make sure about their final scores. Sometimes, they even needed to compare the papers. 3 of these raters in this opinion were in experienced group (Group 2), 1 of them is in less experienced one (Group 3) and 1 of them is in more experienced one (Group 1). As for the raters who were neutral to the rubric, what was obviously noted that 3 of them didn't use the rubric so they might not be familiar with it. They had different opinions such as no rubric is perfect; experience is more effective than the rubric or they had their own rubric in their mind. These raters took into consideration neither the scoring bands with explanations, nor the additional consideration on the rubric. These raters-T7 and T14- among the more experienced raters (Group 1) and T4 was in experienced raters (Group 2). However, T3 who is in the less experienced raters (Group 3) stated his/her hesitation because of being indecisive in assigning a final score within the band. The rater found no problem in the rubric, but his/her own use of the rubric was not effective as s/he got lost in deciding a score. T6 was also in the less experienced group and the rater's problem with the rubric was the application of 'additional use'.

c) their holistic and analytic scoring preference. Whether instructors score writing exams holistically or analytically and if their preference of holistic and analytical scoring had an effect on the final scores of students' exams were searched firstly by asking instructors' opinions in the questionnaire. Then some of them stated their holistic and analytic scoring preferences and how they scored writing exams in general while scoring the papers and thinking aloud. Finally, during the interviews, they were asked whether they scored the exams holistically and analytically and what the reason is for choosing the way they score.

In the questionnaire, part III, the participants were requested to put a tick to the question 'I prefer to score students' writing holistically or analytically'. 32% of the participants answered they prefer analytic scoring while 68% of them said they prefer holistically.

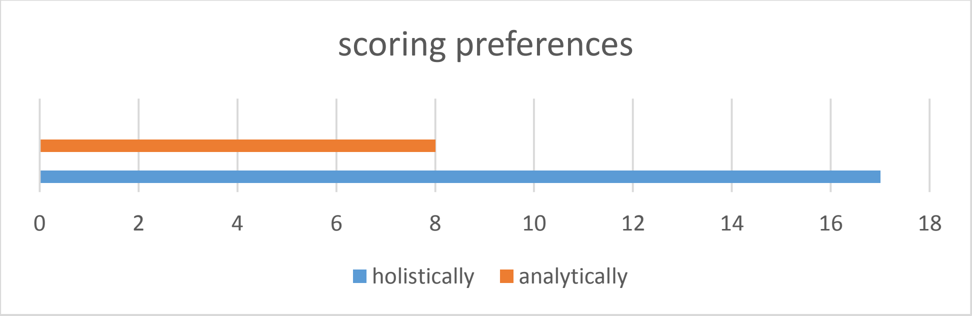


Figure 5. Item 5 of the questionnaire

Later, the participants were asked to share their opinions on which scoring way is more effective than the other. The participants who stated that they prefer holistic scoring also believe it is more effective than analytical scoring and the participants choosing analytical scoring believe just the opposite. As there are various exams in the school like proficiency exam, exemption exam, level achievement exam and progress exams, the participants were asked to give their opinions which scoring way they use in proficiency exams. 28% of the participants refused that they use holistic soring in proficiency exams.60% of the participants agreed that they use holistic scoring. 12% of the participants stated that their use of scoring for proficiency depends on the situations according to whom they score the papers with, or whether it is the proficiency conducted at the end of the year or at the beginning of the year. They stated that for September proficiency, holistic scoring is better, whereas for June proficiency, analytic scoring is better and they do it this way.

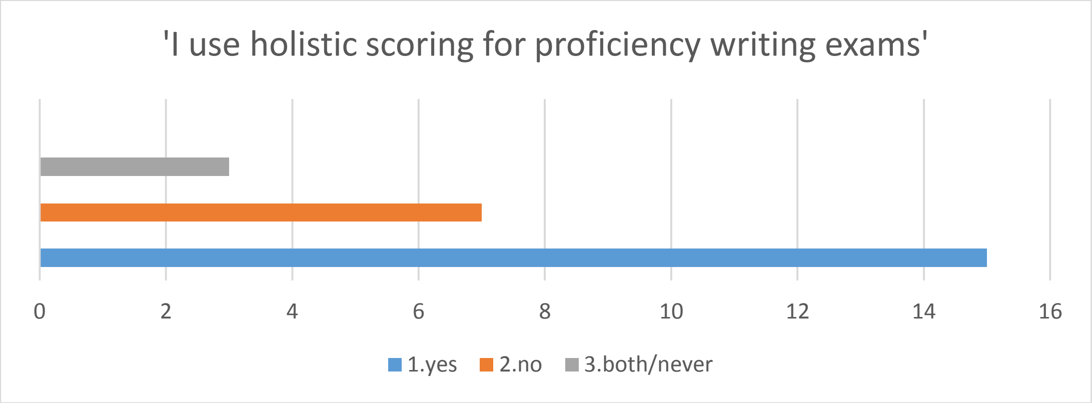


Figure 6. Item 17 of the questionnaire

For other exam types like midterm and final exams, the participants' opinions and use of rubrics were asked in separate questions to make sure. 40% of the participants said that they should score student's midterms exams holistically and 60% of the participants said they should score midterm exams analytically. On the other hand, 64% of the participants believe they should score final exams holistically and 36% of the participants said they should score final exams analytically. Obviously, most of the participants think midterm exams can be assessed analytically while final exams can be assessed holistically when the ratio is considered.

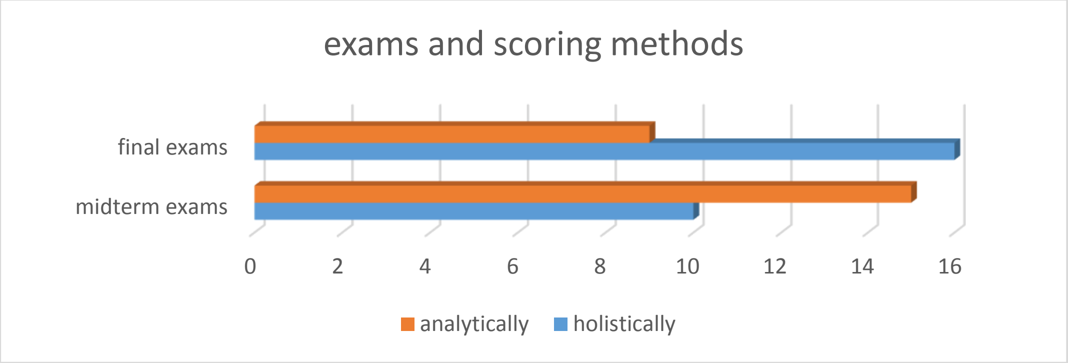


Figure 7. Items 19, 20, 21, 22 of the questionnaire

In order to score student writing exams, instructors were provided with a holistic rubric including 5 scoring bands: good, below average, average, below average and poor. For each band, there are statement guiding the raters in order to have a decision and score the papers considering organization, grammar, use of vocabulary, tone and mechanics, content and how much the paragraph addresses the prompt. Normally, raters weren't warned about their holistic or analytic scoring, neither were they informed about what kind of rubric they were given. To say, they already knew the rubric and used it beforehand, but as they were supposed to know what type of rubric they already used, they weren't explained that the rubric they used was a holistic rubric. While assessing the papers, some of the raters stated how they score the papers:

'I usually assess the papers overall, holistically, but when I hesitate to put a paper into the suitable band like poor, average, below average, etc., I always use the checklist in front of me.' (T9)

'I generally pay attention to content rather than grammar, so I evaluate the papers holistically…' (T4)

During think-aloud sessions, it was clearly observed that some raters were not sure about the difference between holistic scoring and holistic rubric use. As understood from the statement of T9, when raters never used a rubric to score a paper, they called it holistic scoring or as T4 also stated, when they focused on content, not organization or language use, they asserted that they scored the papers holistically. This issue was asked in the interviews and what they shared with the researcher was as follows:

Table 13

Raters' Scoring Ways-Interview

holistically	T1, T3, T4, T5,T9, T13, T14
analytically	T7, T8, T12
both	T2, T6, T10, T11, T15

The first question of the interview was whether they scored students' writing holistically or analytically. 7 raters said holistically, 3 raters said analytically and 5 raters said they used both holistic and analytic scoring. From the answers of the raters, the first issue observed was again the unawareness of the difference between holistic scoring and scoring without a rubric.

> *'..more holistically. Sometimes I took a look at those details on the rubric, but generally speaking I can say that, holistic.' (T14)*

Some raters did the assessment without using a rubric, but thought this way was holistic scoring. Moreover, as T11 stated, when they looked at the criteria to assign a score, scoring became analytically for him or her:

> *'I can say both. Not just holistically or not just analytically. I make use of both, but as a whole I think I am more holistically than analytically. Of course there is a checklist you know. I pay attention to the items on the checklist, so this is also being analytical.' (T11)*

However, there were also raters who were aware that they were provided with a holistic rubric and consequently they did the scoring holistically:

> *'Well, according to the rubric given, it was a holistic rubric I think, yes, so I scored holistically' (T13)*

The second question in the interview was why they chose the way they scored the papers. The reasons they gave varied:

Table 14

Raters' Reasons for Scoring Preferences

Both	Holistically	Analytically
*to be objective	*it was the rubric provided	*students should be assessed according to grammar knowledge, vocabulary, organization
*experience in teaching writing	*it's just a paragraph	*to assess both content and the others
*help to focus on	*content is more important	*all the teachers should score in the same way
	*not focusing on the details much	

The explanations they made for choosing the way of scoring writing exams change. When the reasons of the raters who scored the papers holistically were analyzed, the main reason noted was the rubric itself. They scored holistically because the rubric was holistic. However, it was also observed that raters who believed content is more important than other criteria like organization or grammar preferred scoring holistically as they thought holistic scoring helped them not to focus on details which are grammar, vocabulary use, mechanics, etc. They said they just focused on whether the student could give the message clearly without considering what structures or how various the vocabulary they use. One rater (T14) also stated that if it were an essay, he would prefer to score analytically, but as it is a paragraph, holistic scoring is enough to guide them. The ones who assessed the papers both holistically and analytically conveyed that using both ways makes them objective. They said they trusted their scoring and added that they could make sure much easily in this way. They generally read the papers twice, in the first reading they scored the papers analytically, trying to figure out how effective they use the language and how organized they could state their ideas and in the second reading, they searched for how much they could communicate and express their opinions well. When they stated they did holistic scoring, they analyzed the paper without a rubric and when they asserted that they did analytical scoring, they read the criteria in the rubric provided and tried to match them with the qualities of student papers. As for the raters who assessed the papers analytically, they disclosed that analyzing content, organization, grammar, use of vocabulary, mechanics and other criteria and scoring the papers accordingly, they did analytical scoring. Again they used the same rubric-holistic rubric but by taking every criterion into consideration, they stated they scored analytically. Some raters also added that all raters should do the scoring in the same way, so they scored in this way.

When raters were asked their holistic and analytic scoring preferences and because of the type of the rubric, whether they had any difficulty in scoring during the interviews, some of them stated their views:

> *‘I have difficulty in evaluating in this way because as I’ve mentioned before, I am used to assessing students generally, analytically but it is, you know, a kind of a holistic rubric. For me, it is difficult to assess holistically because when I look at the paper, I’m searching for some categories: what about grammar, what about organization? For example, in the organization part, I want to see some items such as is there any topic sentence, any supporting sentences, supporting details’ (T12-interview).*

As there is not a specific section explaining what to do when the exam paper lacks some criteria in details or to what extent some errors or lacking sentences should be compensated during scoring, raters were unsure about their own scoring and some of them even had to resort to comparisons of papers to get a better result as T1 also stated below:

> ‘With *holistic rubrics, I have problems because I want to make sure to grade them correctly, so I just go back over and over again and read again, the prompts again. I need to match those two, so it is a little bit difficult compared to analytical rubrics.’ (T1-interview)*

Holistic rubrics included bands with general explanations on how to assess papers and give a scoring band involving 4 different points and raters are expected to decide on a suitable score after they made sure of the band that best shows the students' performances. To illustrate, for a poor performance, a rater can give a score from 1 to 4, or for an average performance, a rater needs to determine a score from 9 to 12. Some raters using the holistic rubric provided were observed and also as they explained that they had a hard time to agree on a score although it was easy to define a suitable scoring band for student exams. As T3 stated, it took them a lot of time to make sure.

> *'I have some problems after deciding my part for example, I think that the writing is average, but I can't decide on to give 12 or 9. Generally I spend most of my time thinking about that. The criteria written in each band are enough, but sometimes I cannot be sure whether it is the best score for that writing. I need to go back to other students' papers. I compare the papers to make sure.' (T3-interview)*

In terms of the difference between using a holistic scoring and analytical scoring, some raters shared their experiences and opinions during the interviews and asserted that the score obtained from an analytical scoring could be higher than the one the raters got from a holistic scoring:

> *'…you give sometimes high scores, higher than you do by analytical scoring because when you think analytically, you think of lots of things with very detailed criteria. But holistically you look in general, so sometimes the scores may be higher than the other one.' (T4-interview)*

> *'…sometimes, when I tried to read the papers analytically, the paper got more points than the one I have in my mind in terms of holistic approach.' (T1-interview)*

Research question 3: Do the instructors have difficulty in scoring students' writing? If yes, what kinds of difficulties do they have? To find an answer to this question, all three research tolls were used. First, instructors were asked what they think about this issue on the questionnaire. Then during the think-aloud sessions, they shared their hesitations or difficulties in scoring with reasons, which were also observed and noted. Finally, they were asked if they had any difficulty in scoring during the interviews.

Whether raters had any difficulty in scoring students' writing exams was searched in the questionnaire and 24% of the participants said they had difficulty, whereas 76% of the participants said they didn't have any difficulty in scoring. The reasons for their having difficulty were investigated via the following methods.

Just before the think-aloud sessions, raters were divided into 3 groups according to their years of experience, however they were unaware of this and they worked individually during the scoring process. Group 1 had the most experienced raters-T9, T7, T15, T8, T14-, Group 2 included experienced raters- T1, T2, T4, T12, T13-, Group 3 had less experienced raters-T5, T6, T3, T10, T11.

During the think-aloud sessions, raters had difficulty in assigning the final score in various ways. The main difficulty they encountered is to decide on one single score although they chose a suitable scoring band. To say, they thought that the paper was in the average band, but they were hesitant whether it was 9, 10, 11 or 12.

'*…although there are criteria, rubric given to us, I couldn't feel ready to give a score immediately. I don't feel comfortable. After looking at the other students' writing exams, I want to come back to this paper again…' (T11)*

'After deciding on the band, I really hesitate more to determine my final score within the band…' (T3)

'Just after reading student exam papers, I couldn't make sure about the final score easily, to be honest.' (T11)

Considering these statements, raters cannot be sure enough to score and the criteria may not be enough to make up raters' minds. When raters waver, they need to assess all the papers first and compare them. This difficulty has mostly appeared among the raters who have less experience than the others.

'9 or 12 can be given to this paper in the average band and by being positive, I'll give 12 in a positive manner.' (T14)

Raters not being able to sure about their final score can also change the result. Between 9 and 12, there is a 4-point- difference, which is a huge number. As they can’t decide on a score within a specific band easily, they just determine a grade according to their mood, or positive/negative approach. Moreover, there are also raters who couldn’t decide on a score within the band easily. During the interview, when raters were asked whether they had any difficulties in scoring, one of them mentioned the situation below:

> *‘I have some problems after deciding my part for example, I think that the writing is average, but I can’t decide on to give 12 or 9. Generally I spend most of my time thinking about that’ (T3)*

The rater here stated that s/he had problems in deciding a grade within the same band; whether to give the lowest, highest or a point in the middle of the band. In such cases, the rater added that s/he compared the papers to make sure.

‘If I have difficulty in or am not sure about the score of a paper because of students’ errors, I assess the paper’s content, grammar and vocabulary use because in my mind, to assess the content of student papers, I take the rubric that we used before into consideration. That rubric included content and organization (10 pts.), grammar (6 pts.) and vocabulary use (4 pts.). I always compare the results attained from both assessments…’ (T9)

As seen above, some raters can even use another rubric which may be the one they are more familiar with, and assign a score according to it. This rubric is an extra and other raters don’t use it at that time. Actually, the rubric used additionally is the one that was used by the school beforehand. These raters need a guidance that shows specific points for each criteria.

Raters also state that they can’t start scoring the papers immediately after the papers were provided to them. They may need to analyze some papers quickly first, maybe compare them, then they can get used to the process and score the papers:

‘Now I will score the rest of the papers just after I read each of them as I got used to the assessing process.’ (T11)

This rater had a look at the first 7 papers quickly, then started scoring them from the very beginning, so s/he read some student papers twice to make sure about his/her own scoring.

While scores are given, almost all raters stated that they can't help comparing the student papers in order to make sure, or they used some statements like 'compared to the previous ones', 'similarly', 'this one is better', etc. during the think-aloud sessions.

> *'After reading this paper, I understood that I should change the previous one as 7, not 8.' (T7)*

> *'Let's say 12. Still I'm just thinking if I can give 13, above average. The papers I have marked up to now were mostly average, below average or poor. So this exam paper looks much better. That's why, I can give above average, 13.' (T15)*

When raters compare student papers with each other, their final score may decrease, as in the first example, or increase as in the second one. Sometimes this rise and fall can be more than 1 pts, which also affects the final score the exam.

> *'I have a band for grading in my mind for this, but to make sure, I will read the next student paper. Maybe what I did is wrong, I shouldn't compare them with each other, but I do this because when I read more student papers, I believe a more valid score to the student both by considering the criteria and comparing the papers.' (T11)*

Assessment differences were also observed during the think-aloud sessions. 4 of the raters read all the papers twice to make sure about their scores. After first reading, they put the papers in an order from worse to better or vice versa and started to assess each of them in detail and scored them.

> *'I try to look at the papers overall … and put the papers in an order from the better to the worse. Then I start assessing in detail…in order to be objective, after finishing scoring, I put them in order again according to the grading sequence and look at them quickly… because I am a graduate of teaching department and we were taught that after a weak paper, a better paper can seem much better or after a good paper, a worse paper seems even worse.' (T9)*

They also stated that they could never score a student writing reading it once, so some raters decide on a score just after the first reading while there are raters who feel the need to assess each paper again and again and compare them to come up with a score. The reason for not assessing the papers in the order given was declared by some raters during the interviews:

> *'Some papers are really distracting. They have very bad grammar and vocabulary use, so it is more difficult to focus on those papers. So at the very beginning, to be more focused, I choose a better one to start with for my own goodness. (T10-intreview)*

When writing assessment takes a long time, or the number of papers to be assessed is more than they expect, raters have difficulty in scoring and lose their concentration quickly.

'Is this the last one? No, okay…' (T14)

'How many left?... I lost my concentration, so let me read again.' (T7)

Although the number of papers to be scored was 12, they got bored and as they got bored, they started to assess the papers much more quickly and gave feedback on just some of the criteria.

Research question 4: Is there a significant difference between/among the student grades when their writing exams are scored by more than one instructor? This is the crucial question which also constitutes the start line of the study. Scoring students' papers is conducted by more than one rater in the university where the research was carried out. After the papers are scored by one instructor, they are given to another so that double scoring could be carried out. When both results show a discrepancy of more than 3 points, a third instructor scores the papers, however, his or her decision is the final score of the paper without questioning. Generally, it is hoped that parallelism can be achieved from the results of the first scoring and the second scoring. Keeping this in mind, 15 instructors working in the same university with different educational and professional backgrounds and from different age groups were selected as raters and they were kindly asked to score 12 student papers of proficiency exam as they always do. The raters were provided with the rubric which they already used in similar exams. They assessed each student paper by commenting on the papers aloud and finalized a score. The scores assigned by the raters can be seen in the table below:

Table 15

Raters' Scores

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
T1	14	9	13	13	9	5	9	10	7	7	5	4
T2	7	2	11	13	3	2	4	8	2	7	4	4
T3	9	12	18	10	7	4	8	10	9	6	6	3
T4	10	9	12	13	9	6	6	3	9	10	8	7
T5	8	3	17	15	9	5	5	7	4	9	5	4
T6	9	6	18	16	10	5	14	10	6	8	6	4
T7	11	7	10	8	6	4	9	9	6	5	6	4
T8	12	8	16	13	7	4	8	9	11	7	9	8
T9	8	2	11	3,5	2	1	2	1	4	3	3	1
T10	6	8	14	16	7	4	7	7	4	9	4	5
T11	15	11	15	16	7	7	8	11	9	10	7	7
T12	9	4	11	13	7	3	5	7	8	8	9	4
T13	10	5	12	10	4	1	9	8	8	4	4	1
T14	8	8	12	13	9	6	9	9	6	10	9	5
T15	8	4	10	9	3	3	3	13	5	6	4	3

The data gathered here was analyzed within each group for per paper, by getting the mode, median and mean scores assigned by each rater for per student paper, comparing the results of the hidden groups which were organized before according to their years of experience within the group members and the other groups. The hidden groups are G1, G2 and G3 according to their years of experience. G1 involves T7, T8, T9, T14, T15; G2 involves T1, T2, T4, T12, T13; G3 involves T3, T5, T6, T10, T11. These 15 raters' minimum and maximum scores along with mode, median and mean scores for each writing exam are given in the table below:

Table 16

Each Writing Exam's Mode, Median, Minimum and Maximum Scores and Mean Scores

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P 10	P11	P12
mode	8	8	11,12	13	7	4	9	7,9,10	4,6,9	7,10	4	4
median	9	7	12	13	7	4	8	9	6	7	6	4
Min.	6	2	10	3,5	2	1	2	1	2	3	3	1
Max.	15	12	18	16	10	7	14	13	11	10	9	8
mean	9,6	6,5	13,3	12	6,6	4	7	8,1	6,5	7,2	5,9	4,2

In the questionnaire Part IV, instructors were asked whether they think there are significant differences among EFL instructors' scoring results of the writing exams and what the reasons are for this difference according to them. Just 1 rater disagreed with the idea of difference and 24 of the participants shared their awareness of difference by stating various reasons like:

* focusing too much on grammar rather than content, vice versa
* teachers' different expectations
* experience
* educational background
* teachers' mood and emotions
* personality traits
* scoring holistically and analytically
* instructors' priority: content, organization, grammar
* raters' not paying attention to the rubrics-rubric use or ineffective rubric use
* expectations of the institution
* giving importance to only communication, the message conveyed
* being teachers of different levels
* if they have a child or not (emotional)
* their thinking that they're rewarding students by giving extra points
* some instructors' being unaware of what language teaching and learning is
* lack of institutional philosophy

Research question 5: While grading students' writing, do the instructors pay their attention on a particular criterion more than others? If so, which criterion or criteria do they focus on more, and why? While assessing student written performances, raters may pay special attention to some criteria knowingly or unknowingly. These criteria change from rater to rater, but mainly they focus on only language use, organization, or content. An example of content-focused assessment can be seen below:

> '*Actually while assessing the papers, I generally pay attention to the content rather than grammar. When I do so, I evaluate the papers holistically I think because if the student has good grammar but weak content, it means nothing, but if a student writes a rich content with limited grammar knowledge, grammar shouldn't be a reason for taking off grades I think and should stay in the background' (T4)*

During scoring, some raters declared that they paid attention to grammar mistakes more for midterm exams and the exams that affect their performance during the term, nevertheless, when it comes to proficiency exam, they didn't cut off many points because of grammar errors. Instead, they look for how well students express themselves, to say, a good content:

> *'…if these papers were the ones I assess during the term, I would score them differently. I would take off more grades because of grammar mistakes.' (T8)*

Actually, in this study, raters were observed that they had different judgments on grammar mistakes. While some of them cut off points substantially for grammar mistakes as they thought such mistakes hindered the meaning and made the message difficult to be understood, others believed that grammar errors in proficiency exam are an important indicator of how successful a student is in B1+ level, which is the necessary level to pass the proficiency exam in the school where the study was conducted:

> *'There are really big problems in making sentences to make the message understood' (T15)*

> *'…the student tried to write in an organized way, but grammar mistakes and language errors made it a poor paragraph.' (T14)*
>
> *'The student doesn't have enough grammar knowledge so because of the poor grammar, it is not clear what the student is talking about, so it is 5.' (T12)*
>
> *'The student had basic grammar, vocabulary and word choice errors. These errors hinder the meaning of the message.' (T11)*
>
> *'She has some grammar mistakes which do not change the meaning at all' (T13)*
>
> *'…the errors in grammar, vocabulary, and mechanics are not drastic. I understand what she or she means to say. The student doesn't make too many errors that are not suitable to B1+ level...' (T10)*

As can be understood in the statements of the raters above, raters can be divided into two groups in terms of assessing grammar. Some pay attention to correct use of grammar, but others just look at how any grammatical error changes the meaning.

> *'I don't want to give 16 because the student made mistakes like ''Finally, students must hardworking''. As students who pass the proficiency exams in B1+ level here, we don't expect students to make such a mistake (grammar) in this level, so this is 14.' (T10)*

However, raters might have a tendency to be lenient or severe in scoring considering the students' performance in grammar. Some raters can be more sensitive to grammar usage of students and connect their performance in use of English with the proficiency level, not considering the other sides of the performance like correct use of organization or mechanics or well developed ideas, etc. This situation was also asked to the instructors who participated in the questionnaire. They were requested to put a tick to yes or no boxes for the question '*Although there is a rubric given, I have a tendency to give lower or higher grades considering students' grammar*'. 36% of the participants said no and 64% of the participants said yes. This means more than half of the participants think that they are influenced by students' use of grammar and they could be lenient or severe accordingly. On the contrary, in think- aloud process while raters were scoring papers, just the opposite situation was observed:

> *'…as this is the proficiency exam, I try to figure out whether the students are proficient or not, so the owner of this paper had ideas, but these ideas are very limited… while scoring students writing, I firstly look at whether the student understands the topic and discusses it in a good way, to say, whether the student has a content in the writing. Later, I focus on language use, grammar and vocabulary use.' (T9)*

As can be seen here, this rater is content-focused and for him/her, language use comes second and organization was never mentioned. The rater says '*for writing a good content, the student needs to have at least average level of grammar and vocabulary knowledge*', therefore, content is the most important component of a good writing performance for her.

> *'…so as this is the proficiency exam, we assess the papers considering whether they can express themselves.' (T8)*

The rater just focused on the content of the student to finalize a score. Whether students can develop ideas and express them well are the key factors for these raters. Sometimes, raters focus more on organization:

> *'This paper is also like free writing, unorganized…the student used the language well... the student doesn't know the basic paragraph organization, so it is 4.' (T6)*

> *'…this is not an organized paragraph. Just because of this, I can take off half of 20 points…' (T4)*

> *'Actually, the student's language use is good, but because of organization, I take off lots of points, so it is 9.' (T3)*

For these raters, organization is more important than the other criteria as they cut off many points for this. T12 is also one of these raters as s/he only mentioned and commented on organization and grammar structures of the student performances and never mentioned content on any of the papers. Some other qualities were also observed to be taken into considerations during the assessment process. One of these qualities was whether a title was written or not.

> *'There is no title. I expect students to write a title to help the reader be familiar with the topic.' (T8)*

Although there was not a criterion on whether a title is stated, some raters may expect and look for it. Another criterion not stated clearly on the rubric is about memorized phrases students use during writing exams. On this issue, ideas can change from raters to raters and their considerations also affect the scoring.

> *'…here there are no fixed and memorized structures and the student tried to discuss the opinion at least using his or her own words' (T8)*

> *'…the student memorized some structures and in the first and last sentence, he or she used these structures. As long as they use them correctly, I don't think it is a big problem.' (T1)*
> *'…for example, 'in the light of this information, which have been mentioned above', I haven't seen the same or similar structures in the difficulty being used above.' (T13)*

About memorized structures, one rater tended to cut off some points, whereas another rater ignored them and just focused on correct use of structures. These variances can cause differences. T13 assessed a student's performance by comparing the memorized sentence and the student's own sentence so that s/he scored the paper in this sense.

Some other qualities which are not related to students' achievement in language can also be taken into consideration by raters or can affect raters' scoring positively or negatively. One example was observed as seen below:

> *‘The student had a good, clear handwriting. I like such papers as it is easy to read and assess for me… the paragraph is not organized…the paragraph is irrelevant. My score for this is 8.’ (T8)*

The rater was influenced by the student’s handwriting and assigned a score as 8 although the paper was irrelevant and needed to be scored out of 4. These are the criteria set by the rater’s own, not stated in the rubric and change the result.

> *‘…the paragraph is too short to meet the word limit which is between 150-200.’ (T4)*

Among all the instructors attending to the scoring process, this rater was the only one who mentioned word limit in writing. There is nothing stated about the word limit on the rubric even though such a limit was given in the instruction on student exam papers. This rater was careful about word limit which was warned on exam paper, but there is no guidance about it on the rubric provided.

Some raters have a different style and read the papers twice. In the first reading, they care about language use. In the second reading, they focus on organization and content:

> *‘…first of all, I’m just reading it to have a general idea of it and looking at the grammar and vocabulary, not the content or organization. Later, I focus on them.’ (T2)*

This rater read all the papers twice each time regarding a different criterion, but in sequence. There are also other raters who assess the papers twice, but after the first reading, they put the papers in an order from low performance to higher or vice versa. T11, T10 and T9 were some of these raters. Their assessment styles are different from other raters. They also spent more time scoring the papers compared to those who assessed the papers in the same order given to them.

In different judgments of raters, if gender has a role or not was also searched via SPSS. In order to learn about whether female raters tend to be lenient in scoring or they are severe and cut a significant point, a Mann Whitney test was applied and the results can be seen below:

Table 17

Rater Gender

Gender	N	Mean Rank	Sum of Ranks	U	Median	Z	P
Male	3	10.000	30.00	12000	2	-.868	.385
Female	12	7.50	90.00				
Total	15						

A Mann Whitney U Test was used to question if gender has a role in any scoring difference and it showed an insignificant difference in gender attributions of male assessors (*Median* = 2, *n* = 3) and the female assessors (*Median* = 2, *n* = 12), *U* = 12000, *z* = -.868, *p* =.385, which is less than .01.

So as to see the raters' judgments and assessments for the same paper, a student writing exam was selected and the results including the comments and the scores are shown in the table below:

Table 18

Raters' Scores and Comments on a Selected Writing Exam

Raters	Paper 3	Scores
T1	Some grammar mistakes, spelling mistakes, the paragraph is clear to understand in general, organization is good.	13
T2	It has an organization so you can read it easily, some grammar mistakes, content is poor, grammar and vocabulary knowledge is not very good.	11
T3	The student gave the opinion clearly with supporting sentences, reasons, examples, few grammar errors, some sentences are irrelevant to the unity of the paragraph	18
T4	A good content with examples, reasons and supporting sentences, this paragraph cannot be written any better, but grammar can be improved a little, incomplete topic sent	12
T5	Good topic sentence, the answer is relevant to the topic, no grammar problems, just one vocabulary, one spelling problem.	17
T6	Good organization, majors are good. The student expressed his or her opinion and there is no irrelevant sentence. A few slight grammar mistakes but not important.	18

T7	The student had an opinion to write about, couldn't express what he or she meant exactly.	10
T8	No title, unity between topic sent and concluding sent, some grammar mistakes, examples were good despite being too long	16
T9	I gave the highest score to this, I understood clearly why the student wanted to stay alone, the student mentioned at least 3 reasons and connected them with transitions. There are also grammar mistakes	11
T10	Organization is not problematic, in terms of grammar and use of vocabulary, there are problems which affect meaning, but these mistakes don't hinder the message given.	14
T11	Better at giving the message, supports the topic with more examples, content was good, it has very simple grammar errors	15
T12	A better organization compared to those two, full of grammar mistakes, vocabulary needs to be developed.	11
T13	The paragraph is clear, but it includes more than a few errors in grammar. The paper has examples, details. Although there are a lot of grammar mistakes, she can convey the meaning.	12
T14	More organized compared to first two papers, there are expression problems and spelling mistakes.	12
T15	There are only a few grammatically correct sentences. I can get the message clearly, but examples aren't related to the message.	10

For the same student paper, comments and judgments shared by 15 different rates can be seen in Table 18. 3 rates assigned a score in good band, 4 raters assigned in above average band and 8 raters decided on the average band for this paper. While the highest score is 18, the lowest score is 10 points and there is an 8-point- difference, which makes it necessary to be rescored by another rater. T3, T5 and T6 are the raters who gave the high scores to the paper and they were in the less experienced group (Group 3), whereas T7, T9 and T14 and T15 were among the raters who gave lower scores to the paper and there were in the more experienced groups (Group 1). For the same paper, while T5 stated that there was no grammar problem, T12 said just the opposite and claimed that it was full of grammar mistakes. T7 declared that the student couldn't express what he or she meant exactly, whereas T3 asserted that the student gave the opinion clearly, or T2 found the content poor, but T11 claimed that content was good. In general, all the raters mentioned and commented on organization, content and grammar of the paper and stated that in terms of organization and content, they liked the paper.

Moreover, it was observed that those who said organization or content of the paper was good tended to ignore grammatical problems or mechanical errors.

In the interviews, raters were asked what mistakes they observed on student exam papers, how they assigned their final score and what they paid attention to on student papers most. Firstly, what main problems they encountered during the assessments are illustrated in the table:

Table 19

Raters' Focus in Scoring- Interview

Content	T4
Organization	
Grammar	T6
Organization and Grammar	T15, T14, T13, T12,T10, T7,T5
Content and Organization	
Content and grammar	T9
Errors related to all criteria	T11, T8, T3, T2, T1

Raters generally came across organizational and grammatical errors on student writing exams. The main organizational problem was students' being unable to write a good topic sentence. They believed organizational mistakes affected the content. In terms of grammatical errors, they claimed that students made very simple mistakes, just translated from Turkish to English, wrote words side by side as if they were a sentence. Some memorized structures also distracted attention as compared to other sentences, their language proficiency were easily recognized. Raters generally assessed the papers in terms of grammar within the light of what to be expected from a B1+ level student keeping CEFR levels in mind. As for content, irrelevant sentences, minors not supported enough and the messages that students tried to give being unclear were the main problems they stated. Raters' main focus on a writing exam with a weak performance can be seen in the table below:

Table 20

Raters' Main Focus on a Weak Paper

Content	
Organization	
Grammar	T15, T2
Organization and Grammar	
Content and Organization	T1, T11, T6, T4, T3,
Content and grammar	T14, T12, T9, T8,T7,
Errors related to all criteria	T13, T10, T5,

Raters' judgments and main focuses in assessing low level student performance (P 6) are clearly seen in table 20. It is observed that mostly the focus is more than one criteria. Raters' main focus while scoring all 12 writing exams in general can be seen in the table below:

Table 21

Raters' Observed Focus- Think Aloud Sessions

Content	T11, T9,T7,T4,T2,T15
Organization	T10, T6, T12
Grammar	
Organization and Grammar	T14, T1,
Content and Organization	T8, T3,
Grammar and Content	T13, T10
All the criteria	T5

The most remarkable result attained from the main focus of raters was their not putting grammar into their top priority. Although almost all raters paid attention to grammatical mistakes of students on the papers, they mainly ignored them while scoring. 6 of the raters said that content was their main focus, for 2 of them, it was organization and content, for 3 of them, it was just organization. The raters who paid attention to grammar also paid attention to another criterion like organization and content. There was just 1 rater who stated that s/he focused on every detail to assign the final score.

Conclusion

In order to find answers to 5 research questions, three different methods- a questionnaire, a think-aloud technique, and an interview- were used. Firstly, the participant raters' ages, years of experience and fields of study and gender were compared with tables, charts and SPSS Kruskal Wallis test and Mann Whitney U test apart from getting raters' opinions and listing the reasons. Then raters' utterances and comments during the think-aloud sessions and interviews were stated and their clarifications behind their scoring were compared especially for the same writing exams. Moreover, raters' holistic or analytical scoring preferences for different exams were visualized with tables. Later, the scores raters assigned for the same 12 students were shown and their mode, median, mean, minimum and maximum scores were shown in tables to compare the marks/scores better and see the scoring variance clearly. Finally, some writing exams were chosen and the scores they received were shared with the comments of 15 raters. Additionally, to show raters' focuses and tendencies towards grammar, content, organization, both raters' own explanations and what they did during the assessment were visualized with tables. In discussion part, all these findings will be examined in detail.

Chapter 5
Conclusion, Discussion and Suggestions

Introduction

Acquiring reliable results from an assessment is the ultimate purpose even though it is quite challenging in the assessments which are inclined to subjectivity such as writing and speaking assessments. During writing assessment, no matter how many raters are included in scoring process, unreliable scores can be attained. So as to understand the reasons behind such unreliability, using 3 different techniques, the results have been gathered. In the following part, the findings of the conducted study will be considered and discussed at the hands of the judgments of the researcher in light of literature. Considering research questions and to find each of them answers, all the results will be regarded, and discussed in detail. Not only all the points that can be deduced, but also some suggestions and pedagogical implications will be shared in the end.

Discussion

The role of experience in scoring. In the study, so as to find out if the years of experience in teaching and assessing has any effect on scoring results, firstly the participants of the questionnaire were asked about their years of experience. The number of the raters who have more than 20 years of experience in teaching was more than 50%, which means the raters are quite experienced compared to the number of the raters who have less than 5 years of experience. Moreover, according to the questionnaire results, the highest ratio belongs to the raters with 16-20 years of experience at the university. Considering these, it can be stated that raters at the university where the study was conducted were mostly experienced both in teaching and assessing and they were not new to the system, student profile and administrative goals. They had been through the scoring writing processes many times. The role of novice teachers and experienced teachers in scoring differences in writing and if raters' experience has an effect in this was searched in many studies (Wolfe et al, 2016; Wolfe, 1997; Weigle, 1999; 1998; Pula & Huot, 1993). Therefore, to see the quantitative and qualitative results on this issue, the raters who participated in the think-aloud sessions and follow-up interviews were previously grouped as more experienced (G1, N=5), experienced (G2, N=5) and less experienced (G3, N=5) in terms of their years of experience and their answers in the questionnaire were analyzed via a Kruskall Wallis test. The results showed statistically insignificant group differences within the scope of their years of experience. Similar results were also reached in the study by Attali (2016), which showed that the only difference was inexperienced raters' giving much higher or lower marks to good and bad writing papers although this didn't change the average score compared to experienced raters' marks. As the age may also be considered connected to the experience, participant raters' ages were also analyzed using Kruskall Wallis test and similarly the result showed insignificant difference among groups, which means raters' age and years of experience don't have any effect on scoring differences in writing quantitatively.

To understand better and to learn more on if the raters' educational backgrounds have any influence in scoring ways and scoring results, the effects of their fields were analyzed with a Mann Whitney U test. The fields were divided as ELT and others and in terms of these group attributions, an insignificant result was reached quantitatively. On the other hand, when they were asked about their opinions on whether experience has a role in scoring differences, 68% of the participants stated that experience or lack of experience is a reason for such variance in scoring. They also stated that especially in their early years of teaching, they had a tendency to give higher or lower scores as they focused more on grammar and failed to notice content of student writings. Raters mainly indicated being grammar-focused, too strict, being unable to use rubrics effectively. However, the most outstanding result was giving redundant importance to one criterion: grammar. As stated by some researchers (Attali, 2016; Engelhard, 1994; Meier, 2012; Park, 2008; Schaefer, 2008) these behaviors are called rater effects which include leniency and severity, ineffective rubric use, not even using a rubric at all, or raters' own hesitation in giving a final score the best and worst student performances-halo effect- because of their inner judgments and considerations.

As for the qualitative aspect of the study, the results gathered from different experience groups, age groups and field groups were remarkable. When raters were asked to explain what has changed as they have got more experienced in the questionnaire, they argued that they are more analytical- considering all the aspects of writing-, score holistically- they think they are already familiar with the rubric given and they know the criteria by heart, so they don't need to use the rubric during writing assessments-, focus more on the message students give, not minor mistakes and structural mistakes- they tend to give higher marks to students whose content is rich compared to the students whose grammar is better. However, some raters also declared that they were more careful about the rubric use, especially in some troublesome situations like writing an essay instead of a paragraph, off the topic, etc. Overall, considering their utterances, it can be clearly stated that raters become more content focused and paid attention to the message students communicate instead of being grammar or organization focused as they get more experienced in teaching and scoring. However, in the school where the research was conducted, there were different experienced instructors and hence their marks showed variance as their focus on student writing varied, which caused inter rater unreliability. To make it more clear, for the same student writing, 2 raters' opinions are as follows:

> *'I normally don't care about whether the student wrote a paragraph or an essay as it is a proficiency exam. Students may ignore it, forget it, or are not careful about it.' (T9)*
>
> *'…it has multiple paragraphs, not a single paragraph. For this, maximum score is given as 12 in the rubric.' (T6)*

In this example, T9 is a more experienced rater (G1) and T6 is a less experienced rater (G3). These different experienced raters have different opinions on the same student performance, which resulted in scoring variance in the end. Experienced raters think that in proficiency writing exams, no points should be cut off because of the format, organization. Instead, the focus should be on how the message is communicated. However, inexperienced raters don't have such a distinction as they score papers. This problem is a result of intra rater unreliability which also causes inter rater unreliability accordingly.

Raters' use of rubric and style of scoring. In performance assessments like writing, being objective and achieving the best result is quite critical. For this purpose, using a scoring rubric is discussed to be necessary to help collect all the raters under the same umbrella of criteria and direct them throughout the whole assessment process both for the sake of fair judgment and easy conduction. To increase the quality of assessment (Hafner & Hafner, 2003; Jonsson & Svingby, 2007), rubric use is the most widespread practice as it can be adapted and developed considering the purpose of the exam, curriculum issues and administrative purposes (Brown & Abeywickrama, 2010; Tierney & Simon, 2004). In the light of these, to understand whether raters are aware of the importance of rubric use, they were firstly asked in the questionnaire and 92% of them agreed that rubric use is very important in assessing writing, whereas 8% of the participants declared that the necessity of rubric use is about to the level of students. For students showing good or low performance, they mean they don't need to use a rubric. Even some raters who support rubric use stated that rubric is just a guide and needs to give general information and they added that raters need to focus on content more even though the rubric they use includes different criteria to be considered. This proves that although raters think that rubric use is necessary, they don't obey to all the criteria in the rubric, which results in inefficient use of rubrics and consequently different scoring results. As a follow up question, it was asked to what extend raters/scorers need to abide by the rubric, and 86% of them said all the raters should apply it, whereas 12% declared that using a rubric is bound to rater experience. In these answers, it can be said that although most raters are aware of the need for rubric use and they should use the rubric given, some raters are not in the same opinion especially when it comes to application. Once raters were asked to consider themselves out of 10 in terms of effective use of rubrics, the prominent score was 8 out of 10, which seems a high ration even though in practice, the opposite was observed. To give an example, the rubric supplied to them has 'additional considerations' part which directs the raters in giving maximum scores to some unexpected situations. For example, if the response is irrelevant, it needs to be evaluated out of 4, or if the response is given in multiple paragraphs, it is to be scored out of 12. How some instructors applied these rules can be seen:

> *'…the student didn't write a single paragraph,…so I can score it as 10 out of 20' (T3)*
>
> *'…he wrote multiple paragraphs, so I will assess it out of 16.' (T1)*

> *'…the student wrote in an essay format, but I don't care this a lot. Actually by thinking that all raters cut off some points for this, I can also cut off 1 or 2 points.' (T9)*
>
> *'… the student needed to write one paragraph, but he or she wrote 3 separate paragraphs. But as this is a proficiency exam, I'm against taking off points for this reason. But of course some points should be taken off.' (T8)*

As understood from the raters' utterances below, there is an ineffective use of rubric. Raters assess the same essay format student writing in different ways. While some score them out of 20 by cutting off no points, others just cut off 1-2 or 4 points and some of them stated that some points should be cut off. If we consider the video recording effect and assume that they say they will cut off some points just because of their being recorded, the striking variance of their rubric use can be stated as a reason for scoring differences in writing exams.

The same situation also happened in other ways. If students don't give a relevant response to the topic given, their performances are to be scored out of 4. For this criterion, what some raters say are as follows:

> *'…the student totally misunderstood the topic…the paragraph is irrelevant. The student just mentioned and praised his or her family, so it could be 6 or 7. The student never mentioned staying alone or with friends. My score is 8.' (T8)*
>
> *'…this paper is totally irrelevant response, so it is 4.' (T5)*
>
> *'This paragraph is out of topic and is it a multiple paragraph. The student has her point, but there are problems in grammar, punctuation, so it is going to be 1.'(T13)*

The example statements above show that some raters assign a score without checking the rubric, whereas some raters just give 4 out of 4 without assessing it out of 4. On the other hand, there are raters who take this criterion into consideration and assess the same paper as 1 out of 4 considering also the other criteria. Because of different implementations of the rubric, the same student can get 1 or 8 for the same performance, which is a reason of scoring variance.

In order to avoid subjectivity in scoring, many studies were conducted and some suggestions were given such as averaging marks given by different raters, asking an expert in case of a disagreement, averaging the marks of the expert rater and the one close to his/her, and discussing the marks all together (Johnsson et al, 2001) and averaging the marks for the best score is the most common application (Shavelson & Webb, 1991). As some raters get more rubric focused as they get experienced, some other raters become more independent and don't even need to have a look at the rubric supplied to them. This situation might also result in scoring difference. Sample statements below demonstrate it clearly:

> *'I usually assess the papers overall, holistically, but when I hesitate to put a paper into the suitable band like poor, average, below average, etc., I always use the checklist in front of me.' (T9)*
>
> *'…so we should fully focus on the criteria provided with.' (T13)*

One of the raters above is more experienced while the other is less experienced. During the interviews, the experienced rater stated that s/he didn't use a rubric and assess the writing exams overall-which s/he calls holistically- ,yet when s/he got confused, then s/he used the rubric, whereas the less experienced rater declared that s/he focused on the rubric all the time. These two teachers' scores were given previously in Table 15 in the Findings section. Considering their situation, when it is supposed that for the same student, one rater who doesn't use a rubric in scoring and another rater who focuses on each criterion strictly come together to give a final mark and one rater gives 1 and the other rater gives 8 to the same student, the average score makes 4,5. This situation can be encountered if T9 and T13 come together to score writing exams. However, when two raters who have the same styles in writing assessment come together to score the same paper and they give 8 and 6 for the same paper, the same student's final mark will be 7. There comes out a 3-point-difference and such a difference can change a student's proficiency success as fail or pass (Johnson et al, 2001; Kayapınar, 2014; Meier, 2012). As stated, 3-point- difference can result in a student's failure in some exams. (White, 1985).

The Awareness of the Rubric Content. Raters' familiarity with the rubric depends on how often they used it to assess writing. In this sense, during the study, it was found out that there were two rater groups not aware of the content of the rubric provided: raters who are new at the School of Foreign Languages consequently don't have enough practice of it, and raters, especially more experienced ones, who don't use the rubric given while scoring writing. A rater in less experienced group (G3) stated that s/he needed to read and understand the criteria first before starting scoring the papers. This rater also couldn't make sense of the criteria and explanations on the rubric.

> *'I will evaluate the writing criteria because I am not familiar with this. First I will read the rubric and then score the papers… A few errors? How do I decide on a few errors?' (T5)*

Some of the raters who are experienced in assessing writing wanted to read the criteria to help guide them better even though they are familiar with it.

> *'I'll first read the rubric. I'm familiar with it.' (T1)*

The rubric used to score writing exams is a holistic rubric and such rubrics are quite effective especially if they are used by experienced raters as they could finish scoring in a short period of time (Wang, 2009; Yen, 2016). However, what was shared by some raters was that holistic rubric use takes more time than analytic rubric use because they became indecisive in giving a final score after deciding on a band, so they needed to reread the papers and criteria until they are sure:

> 'With *holistic rubrics, I have problems because I want to make sure to grade them correctly, so I just go back over and over again and read again, the prompts again. I need to match those two, so it is a little bit difficult compared to analytical rubrics.' (T1-interview)*

> *'I have some problems after deciding my part for example, I think that the writing is average, but I can't decide on to give 12 or 9. Generally I spend most of my*

> *time thinking about that. The criteria written in each band are enough, but sometimes I cannot be sure whether it is the best score for that writing. I need to go back to other students' papers. I compare the papers to make sure.' (T3-interview)*

Considering the statements, it's clear that the rubric is not enough to guide some raters well, which causes unreliable scores. What was also found out that some raters couldn't score writing exams because of its type and content. The rubric –holistic- had been used for a few years in the school, and the previous rubric included separate sections for organization, content, vocabulary, grammar, etc. They were used to the previous rubric, so they couldn't adapt themselves to the new rubric. This situation may lead to scoring variance among raters especially more experienced ones. An example was noted during the interview clearly:

> *'Controlling idea or ideas given in the prompt not mentioned?' I couldn't get a clue what they mean.' (T9-interview)*

> *'I am used to assessing students separately, according to each category. Later I collect all the scores.' (T 12-interview)*

> *'I needed some separate items for structure, grammar, maybe content.' (T5-interview)*

This obscurity of the clarifications on the rubric may either lead raters to assign scores on which they can't make sure, or not to use it to score writing exams. No matter what the result is, as the score assigned by the raters using the rubric effectively and ineffectively or using no rubric can be different, which results in scoring variance.

Raters' having no or not enough control of the rubric can cause inconsistency in their own scoring. This is also a reason for intra rater reliability. The rater assigns a score, 7 to two different writing papers with no equal qualities and the writing papers were also compared by the rater. Such comparisons can change scoring results (Goodwin, 2016). Or another rater gives variant scores to different papers by thinking similar reasons:

'…it is a free writing, so I will assess it in below average band.' (T6)

'This paper is also like free writing, unorganized…the student used the language well... the student doesn't know the basic paragraph organization, so it is 4.' (T6)

As seen above, different students wrote free writing- which means unorganized, not in paragraph format, but one of them was considered in Poor section (1-4), whereas the other one was considered in Below Average section (5-8). This intra-rater inconsistency because of unfamiliarity with the rubric is also a cause of scoring difference. Also, with the same reason, another rater's score is 9- Average section.

'…the student wrote in a free style. Additionally, there are lots of spelling and grammar mistakes, so my score is 9.' (T4)

These examples make it clear that different judgments and scorings show both intra-rater and inter rater inconsistency. When raters are unaware of what is included or not stated in the rubric or the maximum score to assign, they may give extra marks or cut off unnecessary marks, so this influences final marks of students:

'Okay there are spelling errors, but I think there is no need to take off any points because of spelling errors.' (T6)

Normally, the rubric included criterion on tone and mechanics and some explanations about them. However, when the raters are not mindful of each criterion, their scores can change and this brings about inter rater inconsistency as other raters can be careful about all criteria. According to the raters' views, using an already prepared rubric (64%) is more important than developing their own rubric (36%). As understood from the findings of the questionnaire, raters' opinions show variety in terms of rubric use. While some raters find the rubric they use is enough to guide them sufficiently, some raters preferred to use an analytical rubric as general statements are believed not to guide them throughout the paragraph since they need point by point explanations for each category like grammar, organization or content. These raters stated that it takes more time to assess the writing exams with a holistic rubric, moreover, they need to read the same writing and the rubric again and again to make sure about their final marks. These findings show that when raters are not satisfied with the rubric provided or cannot adapt themselves to it because of many factors like their previous rubric habits, different scoring styles-detailed, overall; or with- without rubric, they can reach at variant scores for the same writing exams. Raters' different scoring styles (Vaughan, 1991) and their different concentrations and changing focuses (Eckes, 2008) cause unreliability in scores.

Raters are mainly observed not to be aware of the criteria stated at the end of the rubric as 'additional considerations'. They are either unaware of the maximum scores to be given to students writing exams with no controlling ideas, in no paragraph format, etc. or ignore such criteria. Consequently, this situation brings about scoring difference. When the following explanations of raters are examined, what is meant will be clearer:

> *'… the student needed to write one paragraph, but he or she wrote 3 separate paragraphs. But as this is a proficiency exam, I'm against taking off points for this reason. But of course some points should be taken off.' (T8)*
>
> *'…it has multiple paragraphs, not a single paragraph. For this, maximum score is given as 12 in the rubric.' (T6)*
>
> *'…he wrote multiple paragraphs, so I will assess it out of 16.' (T1)*

T8 is in more experienced group (G3), T1 is in experienced group (G2) and T6 is in less experienced group (G3). For a writing exam not written in a single paragraph as expected, one rater-T8- is against taking off any points and talks ambiguously when it comes to the score to be cut off, another rater-T1- is not sure how many points are to be cut off according to the rubric and this is because of ignorance, the other rater- T6- is strict to the rubric and assess the paper accordingly. Normally if a student writes in a multiple paragraph format, his or her paper needs to be assessed out of 12 according to the rubric. Because of 3 different scoring ways, different scores are achieved and this causes inter rater unreliability.

For this performance (P1), T8 assigns 12, T6 assigns 9 and T1 assigns 14. As it is seen, P1 can get 9 or 14 from different raters, which makes a 5-point- difference in the final score. This example is also an obvious reason for scoring variance and inter rater reliability.

Another misuse of rubric can be seen in the example below:

> *'the student didn't use various vocabulary items, he or she had grammar and spelling mistakes and there were some incomplete sentences. My score for this is 12. If the student had written in a paragraph format, the score would have been higher.' (T8)*

In this example, the rater is unaware that 12 is the maximum score, not the score they will give for all papers written in multiple paragraph format. They are expected to make an assessment out of 12, but the rater here just marks the paper 12 out of 12 although s/he mentions some other problems and needs to cut off points because of them. A similar misunderstanding also happened in assessing papers with a totally irrelevant response:

> *'…this paper is totally irrelevant response, so it is 4.' (T5)*

> *'…the student totally misunderstood the topic…the paragraph is irrelevant. The student just mentioned and praised his or her family, so it could be 6 or 7. The student never mentioned staying alone or with friends. My score is 8.' (T8)*

> *'This paragraph is out of topic and is it a multiple paragraph. The student has her point, but there are problems in grammar, punctuation, so it is going to be 1.'(T13)*

T8 is in more experienced group (G1), T13 is in experienced group (G2) and T5 is in less experienced group (G3). One rater assigns 4 out of 4 not taking other problems into consideration, another rater assigns 8 although it needs to be assessed out of 4 and the other rater assigns 1 out of 4 considering all the criteria. This paper (P12) gets 1 and 8 at the same time, which makes a 7-point-difference in the final score.

Raters' scoring preferences. In order to understand whether raters' scoring preference, the way they score-holistically or analytically- changes scoring results or causes rater inconsistency, raters were asked about their scoring preference firstly in the questionnaire, then in the interviews. Raters mostly (68%) prefer holistic scoring in writing assessments. 60% of the raters also state that in proficiency exams, they score writing holistically. Actually, some raters (28%) believe they do analytical scoring using a holistic rubric, which means they score every criterion stated in the holistic rubric to make their final judgments compared to those who do not take every criterion into consideration and just do an overall scoring. What is more important is some raters' (12%) having different scoring styles and preferences in different exams. During a term, there are various exams in the school like exemption exam, level achievement exams, progress exams, proficiency exams which are conducted in January, June and September. They state that for September proficiency, they do holistic scoring, whereas for June proficiency, they prefer analytic scoring. Such divergent attitudes in scoring inevitably results in variance in the scores assigned as there is no unity among raters in assigning a score. Although there appear similar results from different conducts, this cannot be said to be away from coincidence. For different exams, raters' opinions were asked for by follow-up questions in the questionnaire. Most raters (60%) believe for midterm exams, analytical scoring is supposed to be used while for final exams, raters mainly (64%) think a holistic scoring will be better. As again there is no agreement, consequently different conducts bear different scores.

When it comes to application, some results are noteworthy. Raters were given the holistic rubric which had been used many times by themselves without explaining it's a holistic rubric. It was observed that some raters are unaware how they score writing:

'I usually assess the papers overall, holistically, but when I hesitate to put a paper into the suitable band like poor, average, below average, etc., I always use the checklist in front of me.' (T9)

'I generally pay attention to content rather than grammar, so I evaluate the papers holistically…' (T4)

As seen above, some raters believe assessing with no criteria is holistic scoring while assessing with a checklist is analytical scoring, which can be considered as a factor in rater-rater and intra-rater reliability in scoring. Raters' marking some writings with a rubric as they hesitate, but not marking others with a rubric as they feel confident causes intra-rater unreliability. If raters assign a mark without using a scoring rubric or just based on a single criterion like content or organization, inconsistent results reveal within the raters' own scoring and among different raters' scores. Supposing two writing performances both take 10 points because of content, but one of them is better organized or with strong vocabulary variety and correct use, it cannot be claimed that they are the same in scores. Such circumstances cause both inter-rater and intra-rater unreliability.

During the interviews, raters were also asked how they did scoring and 7 of them said holistically, 3 of them said analytically and 5 of them said they did both holistic and analytical scoring. Raters' scoring style difference and different understandings also show up in the interviews:

> *'I can say both. Not just holistically or not just analytically. I make use of both, but as a whole I think I am more holistically than analytically. Of course there is a checklist you know. I pay attention to the items on the checklist, so this is also being analytical.' (T11)*

When the statement above is considered, some raters are confused in the way they score or consciously ignore the rubric in some papers and use the rubric effectively when in hesitation. Once raters were asked the reasons behind their scoring preference in the interview, the raters preferring holistic scoring stated that as it is a paragraph, there is no need for details and content is more important than other criteria, whereas the raters preferring analytical scoring claim that considering each criterion like grammar, organization, content, etc. is analytical scoring and each rater needs to assess in the same way. The raters saying they both use analytic and holistic scoring believe they are more objective and they assert that experience plays a key role in deciding where to use holistic scoring or analytical scoring.

Raters' scoring preference has an effect in students' final marks and some raters are aware of this fact. When they choose one way to score, their final scores can be higher or lower:

> *'…you give sometimes high scores, higher than you do by analytical scoring because when you think analytically, you think of lots of things with very detailed criteria. But holistically you look in general, so sometimes the scores may be higher than the other one.' (T4-interview)*

> *'…sometimes when I tried to read the papers analytically, the paper got more points than the one I have in my mind in terms of holistic approach.' (T1-interview)*

According to some raters, when they pay attention to each criteria in assessing writing, and giving a score for each component, students' final scores are higher. In analytical rubrics, let's say, organization is evaluated out of 5, content evaluated out of 5, grammar is evaluated out of 5 and vocabulary use, tone and mechanics are evaluated out of 5. The same writing gets higher compared to the score assigned using a holistic rubric. This is mainly because of not focusing on one criterion as raters may skip some criteria and aspects to consider in holistic scoring.

The difficulties raters have in scoring. No matter how experienced raters are, each rater may have difficulty in scoring because of various reasons. In this study, raters were asked whether they encountered any problems during scoring writing both in the questionnaire and the interviews. Initially, raters were asked if they had any difficulty in scoring in the questionnaire and most of them (76%) said they had no difficulty. However, during think-aloud sessions, the main problem that is observed by the researcher is raters' hesitation to define a single score to a writing exam. This hesitation was also shared by the raters aloud. They couldn't make sure which score to assign in a band even though they decide on the band. It takes too much time to decide and the decision making process that lasts more than enough leads raters to variant scores as they can lose their concentration while reading again and again. In this case, they compare writing exams, put them in an order from better to worse or vice versa:

‘After deciding on the band, I really hesitate more to determine my final score within the band…I have some problems after deciding my part for example, I think that the writing is average, but I can’t decide on to give 12 or 9. Generally I spend most of my time thinking about that’ (T3)

‘*…although there are criteria, rubric given to us, I couldn’t feel ready to give a score immediately. I don’t feel comfortable. After looking at the other students’ writing exams, I want to come back to this paper again…’ (T11)*

When some raters can’t decide on a single score in the same band, they take their final decision in the light of their positive or negative mood or feelings:

‘9 or 12 can be given to this paper in the average band and by being positive, I’ll give 12 in a positive manner.’ (T14)

Raters’ these practices- comparing papers, considering emotions are all because of their being unsure about a score in a band. In average band, there are 4 scores from 9 to 12 that can be assigned. Both 9 and 12 shows an average writing performance. As in the example above, just after determining the band as average, with no further considerations, the rater gives 12, the maximum score in the band just to be positive. Including emotions into the scoring process leads to inconsistency among scores as another rater can make a negative judgment and give 9 in the same band. Between 9 and 12, there is a 4-point- difference and such a difference can affect students’ success in exams determining failure and success of students in foreign language proficiency. Another point is that some raters apply a different rubric apart from the one they are required to use when they have difficulty in scoring:

‘If I have difficulty in or am not sure about the score of a paper because of students’ errors, I assess the paper’s content, grammar and vocabulary use because in my mind, to assess the content of student papers, I take the rubric that we used before into consideration. That rubric included content and organization (10 pts.), grammar (6 pts.) and vocabulary use (4 pts.). I always compare the results attained from both assessments…’ (T9)

In this example, the rater uses the previous rubric used in the school as s/he is accustomed to it or finds it easier to consider. Assessment differences inevitably cause inter rater unreliability because scores assigned using different rubrics show different results.

Raters' comparisons also affect the final scores. They resort to making comparisons among papers either because they cannot decide a score easily, or they think they are more objective in that way. No matter what the reason is, the scores change, which also influence students' success:

> *'After reading this paper, I understood that I should change the previous one as 7, not 8.' (T7)*
>
> *'Let's say 12. Still I'm just thinking if I can give 13, above average. The papers I have marked up to now were mostly average, below average or poor. So this exam paper looks much better. That's why, I can give above average, 13.' (T15)*
>
> *'I have a band for grading in my mind for this, but to make sure, I will read the next student paper. Maybe what I did is wrong, I shouldn't compare them with each other, but I do this because when I read more student papers, I believe a more valid score to the student both by considering the criteria and comparing the papers.' (T11)*

Raters sometimes raise or lower the score by comparing all the papers and this increase or decrease can be 1 point or more without leaning on criteria and this also causes disagreement among raters, consequently unreliability in scoring. Raters mostly question the accuracy of making comparisons while doing it, one rater asserts that it is beneficial in avoiding giving severe or lenient scores:

> *'I try to look at the papers overall … and put the papers in an order from the better to the worse. Then I start assessing in detail…in order to be objective, after finishing scoring, I put them in order again according to the grading sequence and look at them quickly… because I am a graduate of teaching department and we were taught that after a weak paper, a better paper can*

seem much better or after a good paper, a worse paper seems even worse.' (T9)

It is observed that raters compare the writing exams with the previous or following writing exams, not all of the papers, which anyway affects students' performances as a student's writing score is influenced by another student's performance, not by the criteria in the rubric, which also causes inconsistency and unreliability. As previously declared by some scholars, raters' comparing students' writing exams or written productions and assigning a score accordingly affect the final score of students (Goodwin,2016).

Raters also have difficulty in scoring when they score many writing exams at the same time. Marques and McCall (2005) state that raters' having to score a great amount of writing they can cope with results in unreal scores. In this study, raters scored 12 writing exams and some raters find it hard to assess without losing consideration even though they assess more than 30 writing performances for proficiency exams. This also results in scoring variance and intra-rater inconsistency:

'How many left?... I lost my concentration, so let me read again.' (T7)

So as to find out if raters have any difficulty in scoring because of the rubric they use, raters were asked to consider the rubric they use and some raters stated main issues causing difficulty in scoring. 3 experienced raters (G2), 1 more experienced rater (G1) and 1 less experienced rater (G3) mentioned the difficulties they have in scoring because of the rubric provided. The experienced raters (G2) all stated that they would prefer an analytical rubric to a holistic one as they need separate items for grammar, content, organization, vocabulary use, etc. to assess students' papers more focused, and much faster. Especially in exams which designates students' success and a fail or pass as in proficiency exams, their common idea is that a point by point rubric would be much better. On the other hand, the more experienced rater thinks that some additional considerations are not necessary, so they aren't taken into consideration by raters. Additionally, in the holistic rubric given, criteria explained are not clear enough to guide raters. This issue is also expressed by the less experienced rater claiming that a more detailed rubric is to be provided to get similar results from different raters.

According to interview results, 2 more experienced raters (G1), 1 experienced rater (G2) and two less experienced raters (G3) believe that the rubric provided is enough to guide raters sufficiently and they have no difficulty in scoring because of rubric. More experienced raters and experienced raters think the rubric is clear and guides them efficiently as Yen (2016) also states it takes less time for experienced raters to finalize a score, less experienced raters think that more detailed rubric can mislead or perplex raters by adding that raters shouldn't decide on a score just being bound to rubric. Here, it is obviously seen that some raters don't use the rubric all the time or for each writing exam to decide on a score. When it comes to raters who are neutral once asked for their opinions on the rubric use, 2 more experienced raters (G1), 1 experienced rater (G2) and 2 less experienced raters (G3) are hesitant whether rubric use contributes to making the scoring process easier and getting reliable scores. More experienced raters and experienced rater state they never use a rubric in scoring as they think there is no perfect rubric as they have a rubric in their mind, while less experienced raters think that although it leads raters in general, in some situations they cannot decide on a score within the band and have to compare writing exams with each other to get a score. This is another reason for scoring variance among raters. Raters never using the rubric with common criteria which ought to be considered by all raters for each writing assign a score that is much higher or lower than the score assigned by raters who follow the rubric all the time. Moreover, raters who decide on a band like average (9-12) but cannot decide whether it is 9 or 12 in the same band assign variant and inconsistent scores, which results in scoring difference in the end. An example of this issue is observed during the interview:

> *'I have some problems after deciding my part for example, I think that the writing is average, but I can't decide on to give 12 or 9. Generally I spend most of my time thinking about that. The criteria written in each band are enough, but sometimes I cannot be sure whether it is the best score for that writing. I need to go back to other students' papers. I compare the papers to make sure.' (T3-interview)*

These raters have difficulty especially in considering some statements in the rubric like a few errors, adequate or sufficient and they are unsure how much is adequate or sufficient. They also have the same problem in considering additional considerations stated in the rubric. When a student writes an essay instead of a paragraph, this performance is to be scored out of 12, not 20 according to the rubric. However, the student has no errors in grammar and vocabulary or the content is well written. In such situations, raters' scores are quite variant as some believe it is unfair to assign a score out of 12, others just ignore this criterion and assign a score out of 20 in spite of the rubric use. Supposing that a rater who is unsure of his score and another rater who never uses the rubric in giving a score come together to assign a group of writing exams. In this sense, it is impossible to know how concrete, reliable or valid the scores are.

In interviews, some raters also shared the problems they had while scoring. One rater stated using the holistic rubric provided is difficult for her:

> *'I have difficulty in evaluating in this way because as I've mentioned before, I am used to assessing students generally, analytically but it is, you know, a kind of a holistic rubric. For me, it is difficult to assess holistically because when I look at the paper, I'm searching for some categories: what about grammar, what about organization? For example, in the organization part, I want to see some items such as is there any topic sentence, any supporting sentences, supporting details' (T12-interview).*

Here the rater asserted that being used to separate criteria for every single aspect of writing makes the assessment process difficult for raters when they are required to use a holistic scoring in which they cannot see separate items in details. When they are unsure to what extent errors need to be ignored and assessed as good or above average, they have to read again and again to make sure about their scores:

> 'With *holistic rubrics, I have problems because I want to make sure to grade them correctly, so I just go back over and over again and read again, the prompts again. I need to match those two, so it is a little bit difficult compared to analytical rubrics.' (T1-interview)*

As the rater mentioned in the interview, when raters have a difficulty in assigning a score, they either compare the writing performances or read the criteria many times to make sure in holistic rubric use.

Scoring variance among raters. For the sake of reliability, more than two raters are included in the scoring process so as to achieve objective results (Moscal & Leydens, 2000). In this purpose, it is suggested to have more than one rater to increase inter-rater reliability (Cherry & Meyer, 1993; Kane et al,1999). If one rater decided on the writing score of a student, some possibilities like ignoring some criteria, being affected from their own criteria not written on the rubric or assessing without a rubric can be avoided. Such a score includes subjectivity even if a writing rubric guides raters while scoring. In this current study, 15 raters with different experience years and educational backgrounds participated. They scored 12 writing exams individually, by commenting and giving a final score aloud. The results were visualized in a table. Raters' scores were searched and tried to find if there is a significant difference quantitatively via SPSS Kruskal Wallis test. The marks of previously divided and grouped raters according to their years of experience (G1, G2, G3) were compared and no significant difference has been found out. When it comes to qualitative results, however, raters' comments and their final marks show significant differences.

As previously illustrated in Table 16, for each writing paper, there is a maximum 12,5 (P4), minimum 6- point (P11) difference between raters' final scores. Such a range is quite huge and even changes students' final success. According to White (1985), a difference more than 3 points is excessive between the scores assigned by different raters and this causes inter-rater unreliability. For all writing exams, between maximum and minimum scores, the range is more than 3 points. When P1 is analyzed, the mode is 8, median score is 9 and mean is 9,6. For this writing exam, there are raters who score 6 points, which is 3 points less than the average score and 15 points, which is 6 points more than the average score. Considering the possibility of the maximum scorer and minimum scorer's assessing this writing exam, getting variable, unreliable and inconsistent scores is inevitable. In such situations, Johnson et al (2001) suggests 5 ways of increasing consistency of scores which includes getting an average of scores, asking to an expert, etc.

Having an average score is generally supported as it is believed to increase reliability (Shavelson & Webb, 1991; Johnson et al, 2001).

Instructors stated their opinions after they were asked whether there are significant differences among EFL instructors' scoring of the writing exams and 96% of instructors believe there are differences among scores according to questionnaire results. The reasons behind them vary. Some outstanding reasons are focusing too much on grammar rather than content, vice versa; teachers' different expectations, experience, educational background, teachers' mood and emotions, personality traits, scoring holistically and analytically, instructors' priority: content, organization, grammar; raters' not paying attention to the rubrics-rubric use or ineffective rubric use. Some instructors also state that expectations of the institution, giving importance to only communication, the message conveyed, being teachers of different levels, if they have a child or not (emotional), their thinking that they're rewarding students by giving extra points, some instructors' being unaware of what language teaching and learning is, and lack of institutional philosophy are also reasons behind such differences. However, raters' mainly skip an important point: the expected level of language proficiency is B1+ and students' writing exams need to be assessed according to the outcomes of B1+ level which are clearly stated in CEFR. A B1 level student 'can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence' and B2 level student 'can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesizing and evaluating information and arguments from a number of sources' (Council of Europe, 2001, p.61). A B1+ student is expected to write 'an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options' (p.62). In the light of CEFR descriptors, to determine a student's success and failure in the expected level, raters generally waste time with personal issues and considerations either consciously or unconsciously being far from the awareness of what students can do and not regarding this issue.

Raters' prioritized criteria. In writing assessment, it is crucial to assess every criterion included in the writing checklist or rubric given to raters so as to keep objectivity and hinder scoring variance and unreliability both within and between raters. No matter what rubric is used- holistic or analytical, raters are expected to follow the rubric to attain similar results and for the sake of fairness since writing assessment is prone to subjectivity as it doesn't have concrete answers as in multiple choice exams (Fisher et al, 2002). However, raters may have some priority in their minds and they give more importance to one aspect of writing than other aspects. Content can be valued more than grammar, or organization can be the most essential part of writing and raters especially look for some transition signals or how complex sentences they can write instead of focusing on meaning. These unnoticed considerations are one of the most significant points to be considered as they are obstacles in front of reliable and objective assessments. When the example is examined below, a rater's prioritized criteria is quite easy to notice:

> '*Actually while assessing the papers, I generally pay attention to the content rather than grammar. When I do so, I evaluate the papers holistically I think because if the student has good grammar but weak content, it means nothing, but if a student writes a rich content with limited grammar knowledge, grammar shouldn't be a reason for taking off grades I think and should stay in the background' (T4)*

In this rater's considerations, grammar is unimportant as long as content is detailed and developed well although grammar is among the criteria and should be taken into consideration in all circumstances. Some raters declare that in proficiency exams, grammar should not be high priority and it needs to be considered more in midterm exams:

> *'…if these papers were the ones I assess during the term, I would score them differently. I would take off more grades because of grammar mistakes.' (T8)*

Both in midterm exams and proficiency exams, the same rubric is provided with raters, but some raters tend to focus on one criterion more in one exam and another criterion in other exams. This is also personal judgment damaging inter rater reliability. Raters are observed to have different ideas in whether to cut off and how many points to cut off for grammar errors:

> *'…the errors in grammar, vocabulary, mechanics are not drastic. I understand what she or she means to say. The student doesn't make too many errors that are not suitable to B1+ level...' (T10)*
>
> *'The student doesn't have enough grammar knowledge so because of the poor grammar, it is not clear what the student is talking about, so it is 5.' (T12)*
>
> *'I don't want to give 16 because the student made mistakes like ''Finally, students must hardworking. 'As students who pass the proficiency exams in B1+ level here, we don't expect students to make such a mistake (grammar) in this level, so this is 14.' (T10)*

Raters mismatching opinions in grammar use is another reason for unreliable scores. Their hesitation in this issue brings about inconsistency in scoring. This hesitation is also asked to raters in questionnaire and 64% of them agreed they could be lenient or severe in scoring because of students' grammar use although they use the rubric. Apart from content focused and grammar focused considerations, some raters pay attention to how organized students write and cut off points substantially:

> *'This paper is also like free writing, unorganized…the student used the language well... the student doesn't know the basic paragraph organization, so it is 4.' (T6)*
>
> *'…this is not an organized paragraph. Just because of this, I can take off half of 20 points…' (T4)*
>
> *'Actually, the student's language use is good, but because of organization, I take off lots of points, so it is 9.' (T3)*

Although it is accepted that raters have profound criteria among the criteria given, the score they cut off because of these prioritized criteria can change. As seen above, one rater cuts off half the total point, 10 points just because disorganized writing even though some students express themselves in good grammar and vocabulary. There is not a common score to cut off for organization among raters, which results in unreliability among raters and scoring variance. To show different considerations better, P3 is selected and all raters' scores and opinions are shared in a table (Table 16) in findings. For the same writing exam, one rater (T5) says there is no grammar mistake and the answer is relevant to the topic and assigns 17 points, or another rater (T6) says the student has good organization, good content and just a few slight grammar mistakes , so the score is 18 points, whereas one rater (T2) says there are some grammar mistakes and the content is poor by giving 11 points, or another rater (T9) states she assigned the highest score among the papers she assessed, which is 11 points, by stating that the message is clear and expressed well though there are some grammar errors. One rater (T15) even stated that there are only a few grammatically correct sentences and the message is mainly clear, so the score is 10 points. It is clearly observed that there is no agreement among raters in how correct the language is used and how rich the content is written by the student, consequently the scores vary from 10 to 18 points.

To understand raters' priority more, they were asked what kinds of errors they noticed during scoring in interviews and 7 raters claimed that the errors were mostly in grammar and organization, 5 raters claimed there were all kinds of errors, 1 rater said just content, 1 rater said just grammar and 1 rater said content and grammar. To see clearer, one writing exam (P6) was chosen to find out about raters' main focus in scoring. 5 raters focused on content and organization, 5 raters focused on content and grammar, 2 raters focused on just grammar, 3 raters focused on all criteria. On the other hand, when their focus is judged overall, 6 raters were observed to focus on just content, 3 raters focused on just organization, and other raters focused on either organization and grammar, content and organization or content and grammar. Briefly stated, when raters were asked about common problems on papers, they mostly mentioned organization and grammar while they were observed to focus more on content of student writing exams.

Apart from the criteria stated in the rubric, some raters were also observed to pay attention to students' clear handwriting, word limit-students are expected to write at least 150 words- or title use. To learn if raters' gender has a role in taking such criteria into consideration, a Mann Whitney U Test was used and it is understood that gender has no role in scoring variance and personal considerations.

Conclusion. Writing assessment is a performance assessment in which students' responses to a given topic are judged and scored by raters knowing that there is no concrete answer and consequently it involves subjectivity. In such assessments, to attain one single result is challenging because of many factors like raters' experience, personal traits, the rubric raters use, the criteria they give priority to, expectations from students of B1+ level- the level necessary to pass the proficiency exam, some other criteria not stated in the rubric provided, raters' holistic and analytical scoring preferences, their awareness of the style and content of the rubric and their effective use of rubric, the obstacles or difficulties they encounter during scoring writing. Because of these changing reasons, the scores assigned can be variant, inconsistent and unreliable between raters, even in a rater's own scoring. The answers of raters in questionnaires, their utterances during the think-aloud sessions and interviews were examined and to deeper understand, they were discussed and visualized in this and previous sections. In the following part, what is concluded, some suggestions and pedagogical implications will be shared.

Pedagogical Implications and Suggestions

The results of this study demonstrated that writing assessment is quite difficult in that students' performance, achievements in the target language are evaluated with the help of a writing task which is open-ended, meaning that there is not a key to check the answer one by one. Writing performance includes not only sentences that reflect students' correct use of grammar, vocabulary, tone and mechanics and organization, but also students' creative ideas, supportive details and examples. Therefore, writing is a combination of communicating meaningful ideas and good language performance. It obviously has no single criterion to consider during scoring. To make the scoring process easier and get reliable results, scoring rubrics are used by raters and in this study, a holistic rubric that was already a part of writing assessment process at the school was used by the raters to assess 12 English proficiency writing exams.

Results especially achieved during raters' scoring aloud and interviews showed that raters' scores showed inter-rater unreliability and intra-rater unreliability- in some occasions even though they were provided with a holistic rubric. The reasons behind this are varying, however, most common factors rising to the surface are raters' ineffective or even never use of rubric; some outstanding criteria which are the top priority for raters or ignored by raters- rater effects; their contrasting opinions on what to expect from B1+ level written performance (minimum level for proficiency success at the school); individual or personal criteria not written on the rubric; the unfamiliarity with the style or content of the rubric; raters' analytical or holistic scoring preferences instead of the type of rubric provided with them; raters' years of experience; type of exams-proficiency, exemption, midterm, final exams; comparing exams with each other; the number of exams to be scored, the level raters teach to during the term- expectations of raters teaching to high/low level students during a term; institutional goals for the recent and following term or year.

To deal with the problems causing unreliable scores, benchmarking can be used just before each writing evaluation process to remind or enlighten raters in common aspects and considerations. As East (2009) stated, so as to obtain reliable scores, some sample writing exams can be benchmarked by independent raters to meet raters on a common ground no matter which scoring rubric is used. North, Figueras, Takala, Velherst and Van (2005) stated that coordinators are to select writing performances which will be scored carefully and after the discussions, they should raise awareness of on which scores agreement is reached. The benchmarking sessions need to be organized well.

Moreover, in case a resistance comes out, raters can be observed and guided during writing scoring in terms of how effective they use the rubric given for the sake of getting valid and reliable scores (Goodwin, 2016).

Raters can be informed about the importance of rater effects on scoring results. They can be given feedback after observations to show how different ideas are stated by other raters, how much they affect students' final scores, how important rater agreement is and things to be done to achieve high inter-rater and intra-rater reliability (Wolfe et al, 2016).

It is quite crucial to help raters avoid using much higher or lower scores than deserved, so motivating and leading raters to negotiate on writing performances and their own considerations can be suggested. It is asserted that students writing exams should be assessed considering similar qualities so as to reach an agreement among raters and refrain from subjectivity (Huot, 1990).

To deal with lack of concentration, the number of writing performances to be scored should be limited and each writing exam needs to be assessed separately, by different instructors. When a discrepancy of 3 points comes out, a committee can assess the exams.

Apart from real exams, raters can make practice in scoring real writing performances and receive regular feedback on a long term until similar scores from all raters are achieved. In this sense, rater training is highly suggested. In-service training on writing assessment will be beneficial for both experienced and inexperienced raters.

According to the results gathered from the questionnaire, 80% of the instructors have participated in a professional development organization concerning scoring writing, however, most of these organizations were realized during in-service training which took place at the school during initial years of employment, which means quite a long time has passed since the last organization. To say, giving regular in-service training on writing assessment is highly recommended so as to update raters' practices, increase their awareness not only as a teacher, but also as a scorer, increase interactions of raters by exchanging experiences and opinions, remind raters of what to focus on, refresh raters' perspective, develop a more objective point of view, and keep up with the recent approaches as long as they are practical.

To avoid differences in scores among instructors, considering instructors' viewpoints gathered in the questionnaire, benchmarking, standardization sessions, consultation among raters before or during scoring, defining the objectives clearly for each type of examinations like proficiency exam, final exam or midterm exam; using a rubric which has no gap for any circumstance and is clear, easy to follow and with well-defined categories, giving feedback to instructors after their scorings, encouraging instructors to follow the latest studies to have a common line of vision and sight, organizing seminars and workshops by inviting experts on writing assessment, double scoring or consulting a third instructor and discussing all together on writing exams in collaboration, assessing in groups of 3 or more instructors, and deciding on the style and criteria of the rubric to be used by all instructors together and obeying it rigidly can be considered in educational settings and while educating EFL teacher candidates at universities, these results can be reflected and future English teachers can be provided some guidance and practices on writing assessment in the light of shared information. What is more, institutions having similar problems on writing assessment can make use of the findings and results of this study and reconsider their assessment practices.

To understand what influences rating process and quality, further studies are necessary in order not to miss out any reasons like physical conditions of the environment where raters score writing exams, well trained and not trained raters' scorings, scores of raters' using a rubric which they develop by themselves; scoring time of raters and its' effect on scoring results; raters' fields of study and their effect on scores.

Limitations

As the current study was conducted at a state university, instructors were busy with teaching, preparing materials, etc., so it was somehow demanding to find enough volunteer participants who would attend the study wholeheartedly. The number of participants had to be limited consequently and because of this, the quantitative side of the study couldn't provide significant results. In a study which includes a larger sampling, different results can be revealed.

Another limitation of the study is that in the school where the study was made, CEFR is partly used and considered. The curriculum is designed using the objectives specified to 6 levels of CEFR, yet teaching and testing are not in accordance with them. This situation may cause confusion among raters in what to expect from students' performance and instructors may be unaware of the descriptors for B1+ level written production.

The instruments used in the study were selected to understand raters' own ideas without being influenced, however, the instruments themselves may have an influence on raters' judgments causing lack of concentration, halo effect on raters' scores, etc. Therefore, if different instruments were used, the results could differ.

Conclusion

The primary aim of the study was to find out about any reason behind significant scoring differences among raters as this issue has been one of the mostly discussed and suffered problem in the school where I work and where I worked beforehand. In almost all Schools of Foreign Languages, obtaining reliable results from assessments that are exposed to subjective evaluation is notable even though achieving this is not easy. When the data was analyzed, it was understood that there are many factors affecting scoring results that stem from raters, the rubric used, writing assessment's own feature as having no concrete answer like a multiple choice exam. So as to diminish or terminate factors that are possible to do so, all reasons were expressed, exemplified and visualized in the study and it is aimed to contribute to all English language teachers and administrators, curriculum designers or testers of Foreign Languages' Schools by increasing awareness of such a fact and solutions in the light of my study results.

References

Abeywickrama, P., & Brown,H.D. (2010). Principles of language assessment. In Brown, D.H. (Ed.), *Language Assessment: Principles and classroom practices* (2nd ed., 25-51). Longman

Aims Community College (2018). Using rubrics: Holistic rubric#1. Retrieved from https://www.aims.edu/student/online-writing-lab/resources/using-rubrics

Alanen, R., Huhta, A., & Tarnanen, M. (2010). Designing and assessing L2 writing tasks across CEFR proficiency levels. In I. Bartning, M. Martin. & I. Vedder (Eds.), *Communicative development and linguistic development: Intersections between SLA and language testing research* (pp. 21-56). Eurosla

Altay, İ. F. (2007). Pragmatics of testing. *Journal of Language and Linguistic Studies*, *3*(2), 266-288.

Altay, İ. F. (2008). *A suggested syllabus for advanced writing skills at english language teaching departments.* (Doctoral Dissertation). Hacettepe University, Ankara.

Andrade, H. G. (1997). Understanding rubrics. *Educational leadership*, *54*(4), 14-17.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99-115. https://doi.org/10.1177/0265532215582283

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

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. Language Testing, 12(2), 238-257.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, (2*), 49-58.

Black, P. (2002). Testing: friend or foe?: theory and practice of assessment and testing. Routledge.

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? System, 29(3), 371-383. https://doi.org/10.1016/S0346-251X(01)00025-2

Babin, E. H., & Harrison, K. (1999). Contemporary composition studies: A guide to theorists and terms. Greenwood Publishing Group.

Breland, H. M. (1983). The direct assessment of writing skill: A measurement review. ETS Research Report Series, 1983(2), i-23. https://doi.org/10.1002/j.2330-8516.1983.tb00032.x

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). Assessing writing skill: College Entrance Examination Board research monograph no. 11. New York: The College Entrance Examination Board.

Brookhart, S. M. (1999). The Art and Science of Classroom Assessment. The Missing Part of Pedagogy. ASHE-ERIC Higher Education Report, Volume 27, Number 1. ERIC Clearinghouse on Higher Education, One Dupont Circle, Suite 630, Washington, DC 20036-1183.

Brown, D. H. (1994). *Principles of Language Learning and Teaching*. Englewood Cliffs: New Jersey.

Brown, D. H. (2010). *Language Assessment.: Principles and Classroom Practices*. Longman.

Charney, D. 1984: The validity of using holistic scoring to evaluate writing: A critical overview. Research in the Teaching of English 18, 65–81.

Cherry, R. D., & Meyer, P. R. (1993). Reliability issues in holistic assessment. Validating holistic scoring for writing assessment: Theoretical and empirical foundations, 109-141.

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge, England: Cambridge University Press.

Cumming, A. (1990). Expertise in evaluating second language compositions. Language Testing, 7(1), 31-51. https://doi.org/10.1177/026553229000700104

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. The Modern Language Journal, 86(1), 67-96. https://doi.org/10.1111/1540-4781.00137

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

Daly, J. A., & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement*, *19*(4), 309-316. https://doi.org/10.1111/j.1745-3984.1982.tb00136.x

Davidson, M., Howell, K. W., & Hoekema, P. (2000). Effects of ethnicity and violent content on rubric scores in writing samples. *The Journal of Educational Research*, *93*(6), 367-373. https://doi.org/10.1080/00220670009598731

Demirezen, M. (1993). *From sentence to paragraph structure*. Ankara: Adım.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in judgments of writing ability. *ETS Research Bulletin Series*, *1961*(2), i-93. https://doi.org/10.1002/j.2333-8504.1961.tb00286.x

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing writing*, *14*(2), 88-115. https://doi.org/10.1016/j.asw.2009.04.001

East, M., & Cushing, S. (2016). Innovation in rubric use: Exploring different dimensions. *Assessing Writing*, (30), 1-2. https://doi.org/10.1016/j.asw.2016.09.001

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155–185. https://doi.org/10.1177/0265532207086780

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4* (2), 139–155. https://doi.org/10.1016/1060-3743(95)90004-7

Engelhard Jr, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Engelhard Jr, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and composition program with a many-faceted Rasch model. *ETS Research Report Series*, *2003*(1), i-60. https://doi.org/10.1002/j.2333-8504.2003.tb01893.x

Faigley, L., Cherry, R. D., Jolliffe, D. A., & Skinner, A. M. (1993). *Assessing writers' knowledge and processes of composing*. Ablex Publishing Corporation.

Fisher, R., Brooks, G., & Lewis, M. (Eds.). (2002). *Raising standards in literacy*. Psychology Press.

Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English*, 75-85.

Gunning, T. G. (1998). *Assessing and Correcting Reading and Writing Difficulties*. Order Processing, Allyn and Bacon, PO Box 11071, Des Moines, IA 50336-1071.

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing*, *30*, 21-31. https://doi.org/10.1016/j.asw.2016.07.004

Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International journal of science education*, *25*, 1509–1528. https://doi.org/10.1080/0950069022000038268

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In: L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Norwood, NJ: Ablex.

Hamp-Lyons, L. (2005). What is writing? What is “scholastic aptitude”? What are the consequences? SAT I Writing—a trip down memory lane. *Assessing Writing*, *3*(10), 151-156. https://doi.org/10.1016/j.asw.2005.09.002

Heaton, J. B. (1975). *Writing English language tests: A practical guide for teachers of English as a second or foreign language*. Longman Publishing Group.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980). The influence of context position and scoring method on essay scoring. *Journal of Educational Measurement*, *17*(2), 131-134. https://doi.org/10.1111/j.1745-3984.1980.tb00821.x

Hughes, D. C., Keeling, B., & Tuck, B. F. (1983). Effects of achievement expectations and handwriting quality on scoring essays. *Journal of Educational Measurement*, *20*(1), 65-70. https://doi.org/10.1111/j.1745-3984.1983.tb00190.x

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College composition and communication*, *41*(2), 201-213. https://doi.org/10.2307/358160

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson, & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press, Inc.

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing writing*, 26, 51-66. https://doi.org/10.1016/j.asw.2015.07.002

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, *13*(2), 121-138.

Johnson, R. L., Penny, J., & Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication*, *18*(2), 229-249. https://doi.org/10.1177/0741088301018002003

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, *2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational measurement: issues and practice*, *18*(2), 5-17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Kayapinar, U. (2014). Measuring Essay Assessment: Intra-Rater and Inter-Rater Reliability. *Eurasian Journal of Educational Research*, 57, 113-135.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31. https://doi.org/10.1191/0265532202lt218oa

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*(2), 255-264. https://doi.org/10.1037//0021-9010.86.2.255

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12 (1), 54–71. https://doi.org/10.1177/026553229501200104

Marques, J. F., & McCall, C. (2005). The application of interrater reliability as a solidification instrument in a phenomenological study. *The Qualitative Report*, *10*(3), 439-462.

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

Meier, V. (2012). Evaluating rater and rubric performance on a writing placement exam. *Second Language Studies, 31(1)*, 47-101. http://hdl.handle.net/10125/40721

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. *Performance testing, Cognition and Assessment*, 92-114.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation,* 7(3). Retrieved from (https://pareonline.net/getvn.asp?v=7&n=3&sa=U&ei)

Moskal, B. M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment*, *8*(14), 1-5.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(10), 71-81.

North, B. (2005). The CEFR levels and descriptor scales. In Taylor, L. & Weir, C.J. (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity. Proceedings of the ALTE Berlin Conference* (pp. 21-66). The Edinburgh Building, Cambridge: Cambridge University Press

North, B., Figueras, N., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing*, *22*(3), 261-279. https://doi.org/10.1191/0265532205lt308oa

North, B. (2007). The CEFR illustrative descriptor scales. *The Modern Language Journal*, 91(4), 656-659. https://doi.org/10.1111/j.1540-4781.2007.00627_3.x

Popp, S. E. O., Ryan, J. M., Thompson, M. S., & Behrens, J. T. (2003). Operationalizing the Rubric: The Effect of Benchmark Selection on the Assessed Quality of Writing. *Eric*. Retrieved from https://eric.ed.gov/?id=ED481661

Park, T. (2008). Scoring procedures for assessing writing. *Retrieved* from https://tesol-dev.journals.cdrs.columbia.edu/wp-content/uploads/sites/12/2015/05/3.2-Park-2003.pdf

Perlman, C.C. (2003). Performance assessment: Designing appropriate performance tasks and scoring rubrics. In Wall, J.E. & Walz, G.R. (Eds.), *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators (pp. 497-506).* North Carolina, USA: Eric

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In Williomson, M.M. & Huot, B. A. (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*, (pp. 237-265). Hampton Pr.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18–39. https://doi.org/10.1016/j.asw.2010.01.003

Rublee, M. R. (n.d.). Teaching Analytic Writing through Rubrics. *Analytical Writing Rubric.* Retrieved from https://www.skidmore.edu/assessment/documents/TeachingAnalyticalWritingthroughRubrics.pdf

Sakyi, A. A. (2001). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In: Kunnan, A. J. (Ed.), *19th Language testing research colloquium. Fairness and validation in language assessment* (pp. 131–152). Cambridge, UK: Cambridge University Press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, London: Sage.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), 27-33. https://doi.org/10.1111/j.1540-4781.1992.tb02574.x

Siegert, K. O., & Guo, F. (2009). *Assessing the reliability of GMAT analytical writing assessment*. GMAC Research Report 09-02. McLean, VA: Graduate Management Admission Council.

Silvestri, L., & Oescher, J. (2006). Using Rubrics to Increase the Reliability of Assessment in Health Classes. *International Electronic Journal of Health Education*, *9*, 25-30.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493. https://doi.org/10.1177/0265532208094273

Spandel, V. (2006). In defense of rubrics. *English Journal*, *96*(1), 19-22

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4), 1-19.

Tashakkori, A. & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches* (Vol. 46). Thousands Oaks, CA: Sage.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, *9*(2), 1-10.

Wang, P. (2009). The inter-rater reliability in scoring composition. *English language teaching*, *2*(3), 39-43

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145-178. https://doi.org/10.1016/S1075-2935(00)00010-6

Weigle, S. C. (2002). *Assessing Writing. Cambridge Language Assessment Series.* Cambridge: CUP.

White, E. M. (1985). *Teaching and Assessing Writing: Recent Advances in Understanding, Evaluating, and Improving Student Performance. The Jossey-Bass Higher Education Series*. San Francisco: Jossey- Bass.

Wolcott, W., & Legg, S. M. (1998). *An Overview of Writing Assessment: Theory, Research, and Practice*. National Council of Teachers of English, Urbana, IL

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, 4, 83–106. https://doi.org/10.1016/S1075-2935(97)80006-2

Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing*, *27*, 1-10. https://doi.org/10.1016/j.asw.2015.06.002

Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? In: L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Norwood, NJ: Ablex.

Veal, L. R., & Hudson, S. A. (1983). Direct and indirect measures for large-scale evaluation of writing. *Research in the Teaching of English*, *17*(3), 290-296.

Yen, N. T. Q. (2016). Rater Consistency in Rating L2 Learners' Writing Task. *VNU Journal of Science: Foreign Studies*, *32*(2).

Yuan, M., & Recker, M. (2015). Not all rubrics are equal: A review of rubrics for evaluating the quality of open educational resources. *The International Review of Research in Open and Distributed Learning*, *16*(5), 16-38 https://doi.org/10.19173/irrodl.v16i5.2389

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing*, *27*, 37-53. https://doi.org/10.1016/j.asw.2015.11.001

Zhang, B., Xiao, Y., & Luo, J. (2015). Rater reliability and score discrepancy under holistic and analytic scoring of second language writing. *Language Testing in Asia*, 5,1-9. https://doi.org/10.1186/s40468-015-0014-4

APPENDIX-A: Writing Rubric Used at the School of Foreign Languages

WRITING RUBRIC

GOOD **(20-17)**	The paragraph is exceptional in every way. The text fully answers the prompt. The paragraph is well-organized and all claims are supported with examples or evidence. It begins with a solid introduction that contains a clear and relevant topic sentence, is followed by major and/or minor supporting sentences, and ends with an effective concluding sentence. There are no or few errors in grammar, use of vocabulary, tone and mechanics (spelling and punctuation).
ABOVE AVERAGE **(16-13)**	The paragraph is above adequate in most areas and exceptional in some. In the areas where it is not above adequate, it is still entirely acceptable. The text sufficiently addresses the prompt. The majority of the paragraph is clear, focused and well-detailed, but there may be a few areas requiring further development. While it may contain a few errors with grammar, use of vocabulary, tone and mechanics (spelling and punctuation), these errors are not drastic enough to detract from the overall point being made.

AVERAGE (12-9)	The paragraph is adequate in most areas, but exceptional in none. The text partially addresses the prompt. The paragraph is clear although probably lacking in both control and command. Organization may be a slight problem but errors don’t make it difficult to understand. Supporting sentences provide details but are generally underdeveloped. There may be multiple errors in grammar, use of vocabulary, tone and mechanics (spelling and punctuation), but these errors do not, for the most part, detract from the overall writing.
BELOW AVERAGE (8-5)	The paragraph is lacking in a majority of areas. The text doesn’t adequately address any part of the prompt. The paragraph is not clear and is mostly underdeveloped. It is generally unorganized and unfocused. There are frequent errors in grammar, use of vocabulary, tone and mechanics (spelling and punctuation) that distract from the content being provided.
POOR (4-1)	There are significant problems throughout the paragraph. The paragraph is often lacking and the argument, if there is one, wanders and is unorganized. It shows no understanding of basic paragraph organisation. There are significant errors in grammar, use of vocabulary, tone and mechanics (spelling and punctuation).

ADDITIONAL CONSIDERATIONS

	Maximum grade
no response	0
totally irrelevant response	4
controlling idea/s given in the prompt not mentioned	12
personal opinion not stated	16
multiple paragraph / essay format	12

APPENDIX-B: Consent Form for Interview

Consent to Participate in Interview

The questions below will be asked to you just after you have scored students' writing exams. You are expected to think, comment, and score aloud. Your remarks will be recorded if you are willing. I would like to record this interview to make sure that I remember accurately all the information you provide. I will be interviewing with approximately 15 instructors about how they score students' writing exams. I have developed this consent form so that this collected information will be used in published research as well as in academic presentations as long as you give your permission. The interview doesn't include any questions that disturb you, but you're free to stop answering the questions because of any other factors disturbing you. For this study, Hacettepe University Research Ethics Committee approval has been received. If you have any doubt about any aspect of the survey, or if you would like more information about the study, please do not hesitate to ask me.

Confidentiality

All responses that are collected in this survey will be kept strictly confidential. You are guaranteed that neither you, your remarks nor the name of this university will be identified in any reports of this study.

If you would like to be informed about the result of the survey, please add your contact information here:

1. Did you score students' writing holistically or analytically?
2. What is the reason for choosing the way you score?
3. Do you think the writing rubric given to you to score the writing exam is enough to guide you sufficiently, or have you had any difficulties in scoring because of the rubric?
4. What mistakes have you observed on students writing papers generally?
5. While determining the final score, what have you paid attention to most, in other words, how did you determine your final score?

The purpose of the interview has been explained to me. I agree to participate in the survey and I voluntarily consent to this interview being recorded electronically. I also give permission for all the information I provide can be used in the study.

Name of the interviewee: Date:
E-mail: Signature of the interviewee:

Thank you in advance for your cooperation and participation.

Fatma Merve Uzun
MA Candidate at Hacettepe University
Master of Arts in English Language Teaching
Hacettepe University, School of Foreign Languages
Floor:1 Room:1 Beytepe/ Ankara
0505 585 55 20
fmerveolmez@gmail.com

İsmail Hakkı Mirici
Thesis Supervisor,Professor
Hacettepe University
ELT Department
Beytepe, Çankaya, Ankara
0312-297-8585
hakkimirici@hacettepe.edu.tr

APPENDIX-C: Consent Form for Questionnaire

QUESTIONNARE CONSENT FORM

Dear Participant,

I am a graduate student in the department of English Language Teaching at Hacettepe University. For my master thesis, I am conducting a research on EFL instructors' scoring differences of the writing exams. For this purpose, I have developed this questionnaire. With this questionnaire, it is aimed to gather information about the reasons for the scoring differences in the eyes of the instructors.

This questionnaire should take approximately 15 minutes to complete. The questionnaire doesn't include any questions that disturb you, but you're free to stop answering the questions Feel free to ask anything if something comes to your mind. For this questionnaire, Hacettepe University Research Ethics Committee approval has been received. If you have any doubt about any aspect of the survey, or if you would like more information about the study, please do not hesitate to ask me.

Confidentiality

All responses that are collected in this survey will be kept strictly confidential. You are guaranteed that neither you, your remarks nor the name of this university will be identified in any reports of this study. To this survey, participation is voluntary, so any individual can withdraw the survey at any time.

If you would like to be informed about the result of the survey, please add your contact information here:

I am willing to participate in this survey. Signature: ……………………………..

Date:…………………………………..

Name, Surname:………………………

Email:…………………………………..

Thank you in advance for your cooperation and participation.

Fatma Merve Uzun
MA Candidate at Hacettepe University
Master of Arts in English Language Teaching
Hacettepe University, School of Foreign Languages
Floor:1 Room:1 Beytepe/ Ankara
0505 585 55 20
fmerveolmez@gmail.com

İsmail Hakkı Mirici
Thesis Supervisor,Professor
Hacettepe University
ELT Department
Beytepe, Çankaya, Ankara
0312-297-8585
hakkimirici@hacettepe.edu.tr

APPENDIX-D: Questionnaire Administered to English Instructors of the School of Foreign Languages

Part I: About You

The questions below are about you, your education and your experience in teaching. In responding to the questions, please tick one box for each question.

1.Gender: □ female □ male
2.Age: □ 20-30 □ 31-40 □41-50 □51-65
3.Employment status: □Full time □ Part time
4.Year(s) of experience: □ 0-5 □ 6-10 □11-15
□ 16-20 □ 21-30
5.Year(s) of experience at this university: □ 1-5 □ 6-10 □ 11-15
□ 16-20 □ 21-30
6. Highest qualification in ELT (mark the ones you have)
□ Teaching certificate Yes □ No □
□ Bachelor's degree Field: ____________________
□ Master's degree Field: ____________________
□ Doctorate (PhD) Field: ____________________
□ Other (please specify) _________________________

7. What do you think is your level of English according CEFR?

	A1	A2	B1	B2	C1	C2
Listening						
Reading						
Writing						
Spoken Interaction						
Spoken Production						

Part II: Teaching Writing

For each question below, please tick one box, but if an explanation is necessary, please give as much detail as possible.

8. I teach … hours a week. □12 □13-18 □19-25 □more than 25

9. Do you teach skills □ separately or □ in an integrated way?

10. I spend … hours teaching writing. □ 0 □ 1-3 □ 4-5

11. How do you teach writing?

____

____

____

12. Do you give your students a writing checklist that they can use while writing?

□ yes □ no

13. Do you give feedback to students during and after they write?

□ yes □ no If you put at least one tick, how?

____

____

Part III: Scoring Writing

For each question below, please tick one box, but if an explanation is necessary, please give as much detail as possible.

14. I think using a rubric is crucial in scoring writing:
□ yes □ no

15.I prefer to score students' writing:
□ holistically □ analytically

16. I think holistic scoring is more effective than analytic scoring:
□ yes □ no

17. I use holistic scoring for students' proficiency writing exams.
□ yes □ no

18. I believe every instructor should abide by the rubric given in writing exam while scoring.
□ yes □ no

19. I think instructors should score midterm writing exams analytically.
□ yes □ no

20. I think instructors should score midterm writing exams holistically.
□ yes □ no

21. I think instructors should score final writing exams holistically.
□ yes □ no

22. I think instructors should score final writing exams analytically.
□ yes □ no

23. Are the students' writing exams scored by more than one instructor? If yes, by how many instructors are they scored?
□ yes □ no how many: ________

24. I have difficulty in scoring students' writing exams.

□ yes □ no

25. Although there is a rubric given, I have a tendency to give lower or higher grades considering students' grammar.

□ yes □ no

26. Which one is more effective:

a) □ using an already prepared rubric b) □ developing your own rubric

27. How can you consider yourself out of 10 in terms of using the writing rubric given effectively in an exam?

□ 1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □ 8 □ 9 □ 10

Part IV: Scoring Differences

For each question below, please put tick in one box, but if an explanation is necessary, please give as much detail as possible.

28. I think there are significant differences among EFL instructors' scoring of the writing exams.

□ yes □ no (If you say no, pass item number 31; if yes, answer it.)

29. What do you think are the reasons for the difference in students' writing scores?

____

____

____

30. I think the year of experience an instructor has has a role in scoring differences in writing.

□ yes □ no (If you say no, pass item number 34; if yes, answer it.)

31.What mistakes would you make while scoring students' writing exams in the first year of experience? Can you give a few examples?

____

____

____

32.As you get more experienced, have you changed the way you score students' writing? If yes, How?

____

____

____

Part V: Professional Development

For each question below, please give short answers, but if an explanation is necessary, please give as much detail as possible.

33. Have you ever participated in a professional development organization concerning scoring writing? If yes, when? where?

____

____

____

34. Do you think it is necessary to participate in seminars/ workshops/training etc. regularly on scoring writing? Why?

____

____

35. What do you think is the best way to get rid of the difference in scores among the instructors?

____

____

____

Further Comment
If you have any further comments on **English language instructors' scoring difference of the writing exams,** please add here.

APPENDIX-E: Student Consent Form

STUDENT CONSENT FORM

Dear Participant,

I am a graduate student in the department of English Language Teaching at Hacettepe University. For my master thesis, I am conducting a research on EFL instructors' scoring differences of the writing exams. With this survey, it is aimed to gather information about the reasons for the scoring differences. For this purpose, I have developed this consent form. I would like to take a copy of your writing exam to be re-scored by the instructors. Your name and number will be hided. If you have any doubt about any aspect of the survey, or if you would like more information about the study, please do not hesitate to ask me.

Confidentiality

All responses that are collected in this survey will be kept strictly confidential. You are guaranteed that neither you, nor the name of this university will be identified in any reports of this study. To this survey, participation is voluntary, so any individual has the right not to participate in.

If you would like to be informed about the result of the survey, please add your contact information here:

I allow the researcher to use my writing exam in the survey.

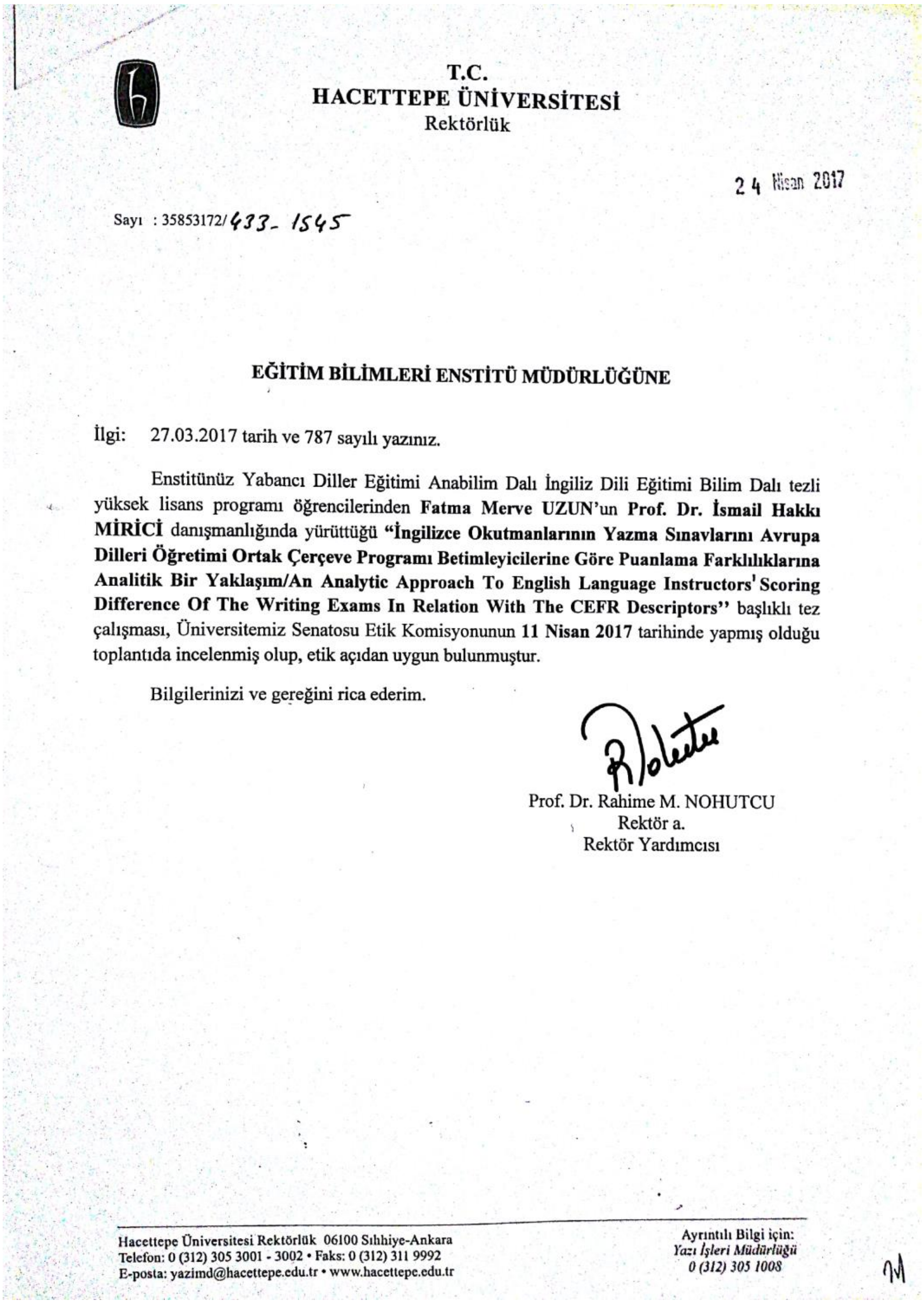
Student name:
Department:
Class:
Signature:

Thank you in advance for your cooperation and participation.

Fatma Merve Uzun
MA Candidate at Hacettepe University
Master of Arts in English Language Teaching
Hacettepe University, School of Foreign Languages
Floor:1 Room:1 Beytepe/ Ankara
0505 585 55 20
fmerveolmez@gmail.com

İsmail Hakkı Mirici
Thesis Supervisor,Professor
Hacettepe University
ELT Department
Beytepe, Çankaya, Ankara
0312-297-8585
hakkimirici@hacettepe.edu.tr

APPENDIX-F: Ethics Committee Approval



T.C.
HACETTEPE ÜNİVERSİTESİ
Rektörlük

24 Nisan 2017

Sayı : 35853172/433- 1545

EĞİTİM BİLİMLERİ ENSTİTÜ MÜDÜRLÜĞÜNE

İlgi: 27.03.2017 tarih ve 787 sayılı yazınız.

Enstitünüz Yabancı Diller Eğitimi Anabilim Dalı İngiliz Dili Eğitimi Bilim Dalı tezli yüksek lisans programı öğrencilerinden **Fatma Merve UZUN**'un **Prof. Dr. İsmail Hakkı MİRİCİ** danışmanlığında yürüttüğü **"İngilizce Okutmanlarının Yazma Sınavlarını Avrupa Dilleri Öğretimi Ortak Çerçeve Programı Betimleyicilerine Göre Puanlama Farklılıklarına Analitik Bir Yaklaşım/An Analytic Approach To English Language Instructors' Scoring Difference Of The Writing Exams In Relation With The CEFR Descriptors''** başlıklı tez çalışması, Üniversitemiz Senatosu Etik Komisyonunun **11 Nisan 2017** tarihinde yapmış olduğu toplantıda incelenmiş olup, etik açıdan uygun bulunmuştur.

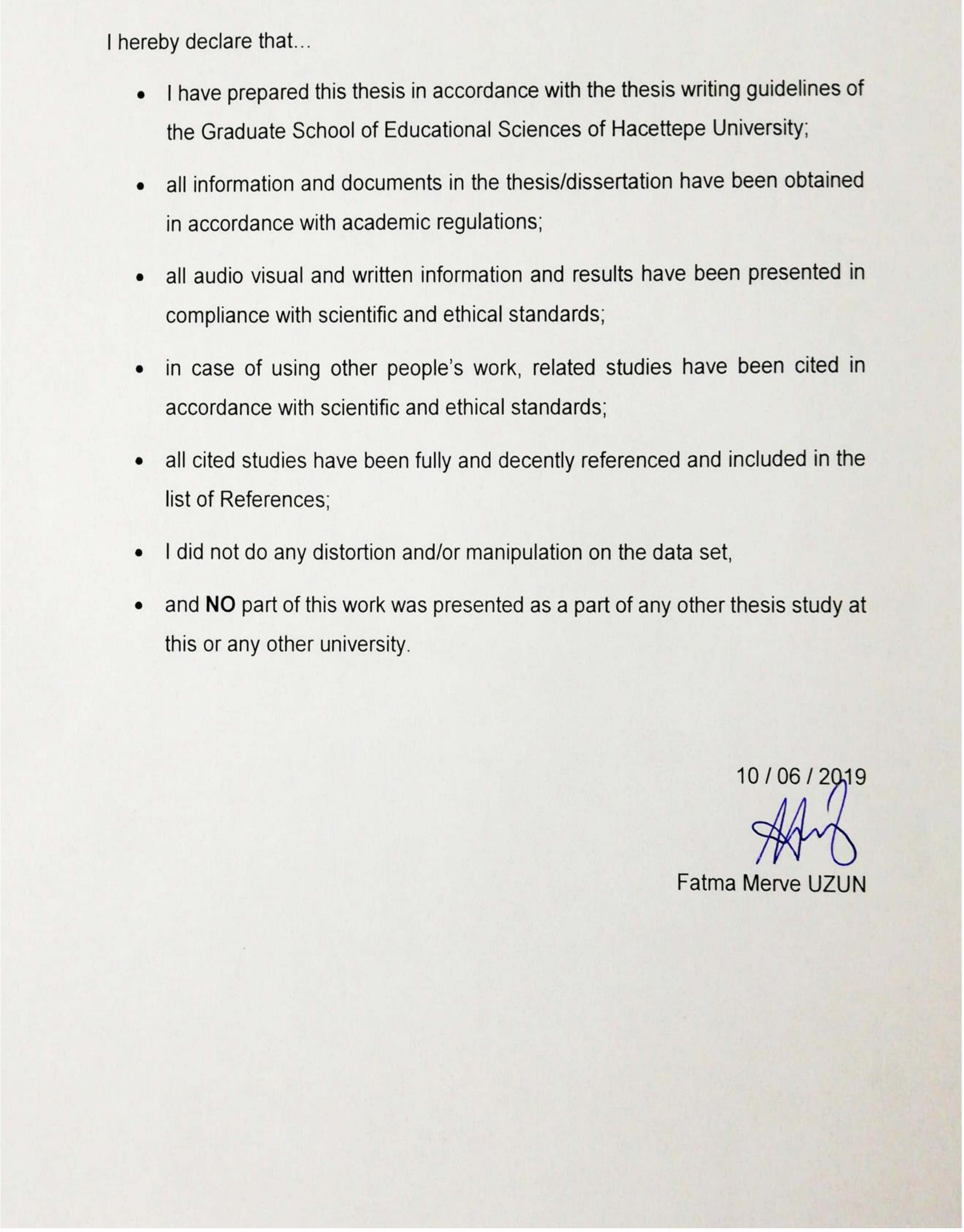
Bilgilerinizi ve gereğini rica ederim.

Prof. Dr. Rahime M. NOHUTCU
Rektör a.
Rektör Yardımcısı

Hacettepe Üniversitesi Rektörlük 06100 Sıhhiye-Ankara
Telefon: 0 (312) 305 3001 - 3002 • Faks: 0 (312) 311 9992
E-posta: yazimd@hacettepe.edu.tr • www.hacettepe.edu.tr

Ayrıntılı Bilgi için:
Yazı İşleri Müdürlüğü
0 (312) 305 1008

APPENDIX-G: Declaration of Ethical Conduct



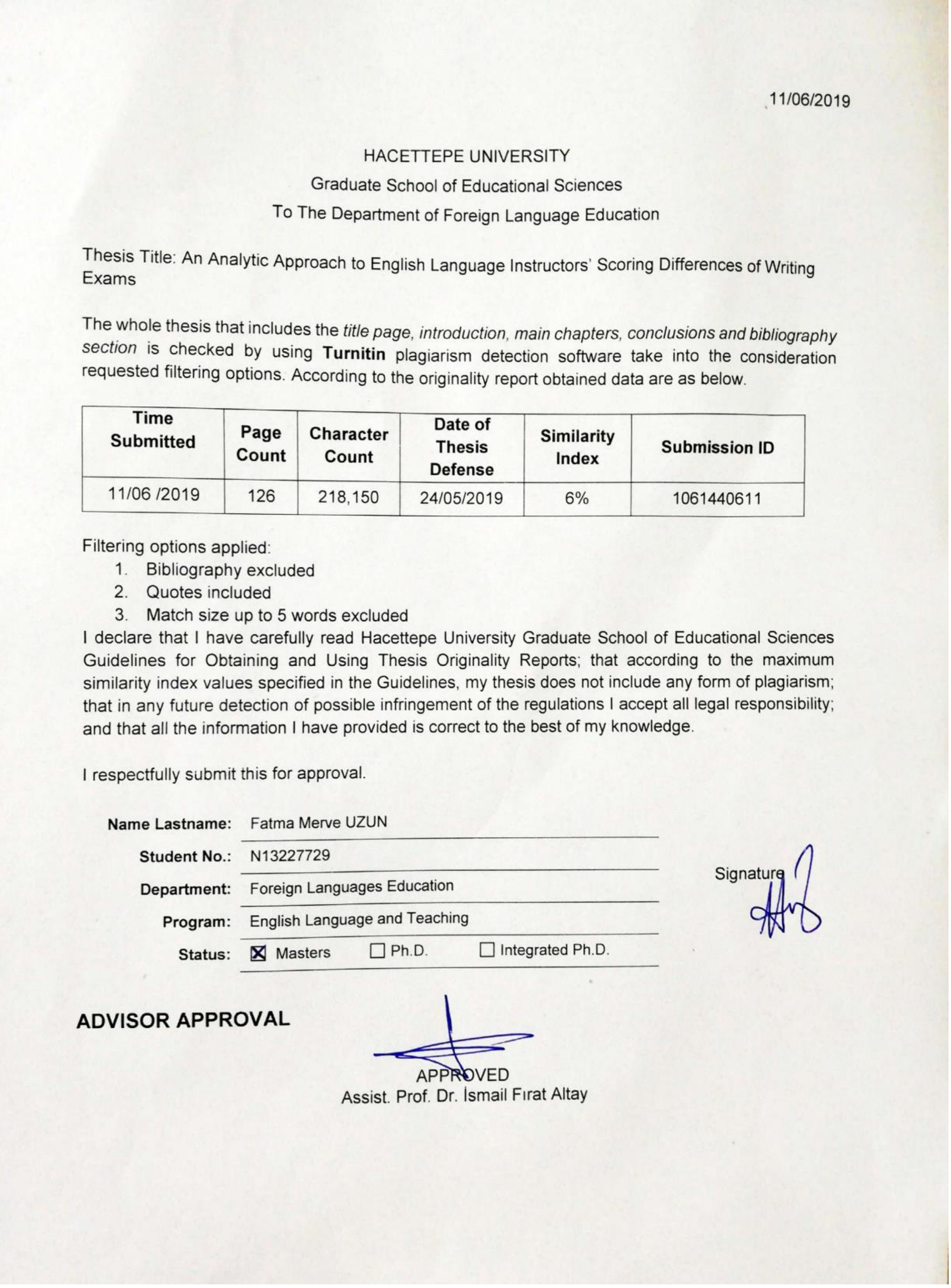
I hereby declare that…

- I have prepared this thesis in accordance with the thesis writing guidelines of the Graduate School of Educational Sciences of Hacettepe University;
- all information and documents in the thesis/dissertation have been obtained in accordance with academic regulations;
- all audio visual and written information and results have been presented in compliance with scientific and ethical standards;
- in case of using other people's work, related studies have been cited in accordance with scientific and ethical standards;
- all cited studies have been fully and decently referenced and included in the list of References;
- I did not do any distortion and/or manipulation on the data set,
- and **NO** part of this work was presented as a part of any other thesis study at this or any other university.

10 / 06 / 2019

Fatma Merve UZUN

APPENDIX-H: Thesis Originality Report



11/06/2019

HACETTEPE UNIVERSITY

Graduate School of Educational Sciences

To The Department of Foreign Language Education

Thesis Title: An Analytic Approach to English Language Instructors' Scoring Differences of Writing Exams

The whole thesis that includes the *title page, introduction, main chapters, conclusions and bibliography section* is checked by using **Turnitin** plagiarism detection software take into the consideration requested filtering options. According to the originality report obtained data are as below.

Time Submitted	Page Count	Character Count	Date of Thesis Defense	Similarity Index	Submission ID
11/06 /2019	126	218,150	24/05/2019	6%	1061440611

Filtering options applied:

1. Bibliography excluded
2. Quotes included
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Educational Sciences Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

Name Lastname: Fatma Merve UZUN

Student No.: N13227729

Department: Foreign Languages Education

Program: English Language and Teaching

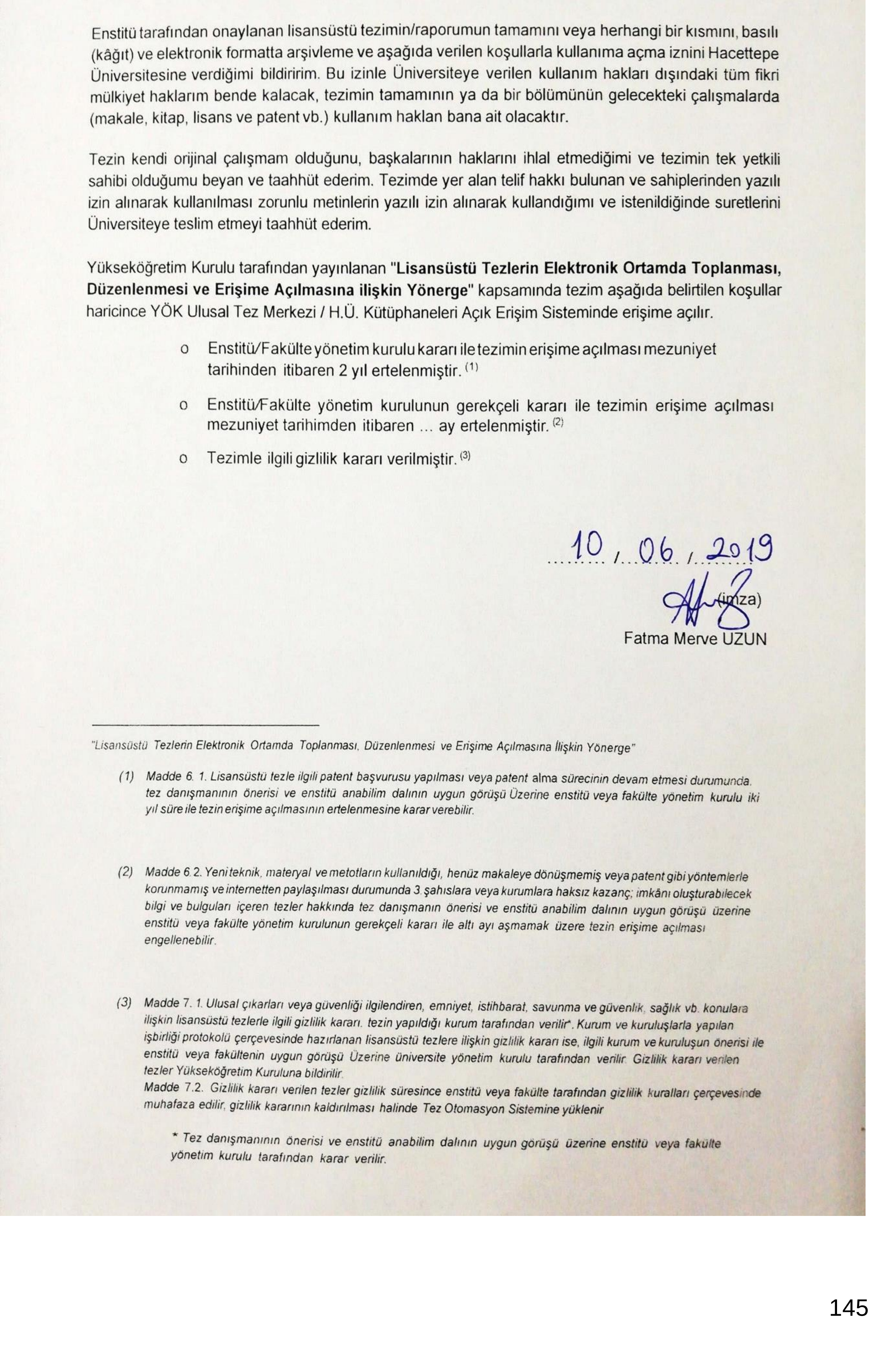
Status: ☒ Masters ☐ Ph.D. ☐ Integrated Ph.D.

Signature

ADVISOR APPROVAL

APPROVED
Assist. Prof. Dr. İsmail Fırat Altay

APPENDIX-I: Yayımlama ve Fikrî Mülkiyet Hakları Beyanı



Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım haklan bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan "**Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına ilişkin Yönerge**" kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

○ Enstitü/Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihinden itibaren 2 yıl ertelenmiştir. (1)

○ Enstitü/Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ... ay ertelenmiştir. (2)

○ Tezimle ilgili gizlilik kararı verilmiştir. (3)

10 / 06 / 2019

(imza)

Fatma Merve UZUN

"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"

(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü Üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.

(2) Madde 6.2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç; imkânı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanın önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.

(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı. tezin yapıldığı kurum tarafından verilir. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü Üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir

** Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir.*