THE REPUBLIC OF TURKEY
ÇANAKKALE ONSEKİZ MART UNIVERSITY
GRADUATE SCHOOL OF EDUCATIONAL SCIENCES
DEPARTMENT OF FOREIGN LANGUAGE EDUCATION
ENGLISH LANGUAGE TEACHING PROGRAMME

THE IMPACT OF RATER EXPERIENCE AND ESSAY QUALITY ON RATER BEHAVIOR AND SCORING

DOCTORAL THESIS

ÖZGÜR ŞAHAN

ÇANAKKALE
FEBRUARY, 2018

The Republic of Turkey
Çanakkale Onsekiz Mart University
Graduate School of Educational Sciences
Department of Foreign Language Educatıon
English Language Teaching Programme

The Impact of Rater Experience and Essay Quality on Rater Behavior and Scoring

Özgür ŞAHAN
(**Doctoral Thesis)**

Supervisor
Assist. Prof. Dr. Salim RAZI

Çanakkale
February, 2018

Declaration

I hereby declare that the Doctoral Dissertation, **“The Impact of Rater Experience and Essay Quality on Rater Behavior and Scoring”**, which I wrote myself, has been prepared in accordance with ethical and scientific values and that all the sources that I have used in this study are included in the references.

05/02/2018

Özgür ŞAHAN

Çanakkale Onsekiz Mart University

Graduate School of Educational Sciences

Certification

We hereby certify that the report prepared by Özgür ŞAHAN and presented to the committee in the thesis defense examination held on 5 February 2018 was found to be satisfactory and has been accepted as a thesis for the degree of Doctor of Philosophy.

Thesis Reference No: 10180258

	Academic Title	Full Name	Signature
Supervisor	Assist. Prof. Dr.	Salim RAZI	....
Member	Prof. Dr.	Dinçay KÖKSAL	....
Member	Assoc. Prof. Dr.	Bülent GÜVEN	....
Member	Assoc. Prof. Dr.	A. Amanda YEŞİLBURSA	....
Member	Assist. Prof. Dr.	Turgay HAN	....

Date:.................................

Signature:..............................

Prof. Dr. Salih Zeki GENÇ

Director, Graduate School of Educational Sciences

Acknowledgement

This dissertation would be impossible without the support of valuable people. First, I would like to express my sincere gratitude to Assist. Prof. Dr. Salim Razı for the guidance and support that he has provided as my supervisor and a valuable friend. He has always enlightened me to be a hardworking and honest scholar throughout my Ph.D. education. Secondly, I would like to express my sincere thanks to Assist. Prof. Dr. Turgay Han for his guidance in developing the methodological and theoretical aspects of this study. He provided me with constructive and helpful criticisms throughout the entire process. In addition, I am very grateful to Prof. Dr. Dinçay Köksal for his valuable support, encouragement, and advice for this thesis. I would like to express my special thanks to Assoc. Prof. Dr. Bülent Güven and Assoc. Prof. Dr. Ayşegül Amanda Yeşilbursa for their guidance and feedback as my jury members.

I owe a thousand thanks to my participants. It would not have been possible to compose this research without their invaluable contributions. I would like to thank my friend Mustafa Çoban for his help and support as a friend and a colleague. My sincere appreciation to my beloved wife Kari for her endless patience, love, and support as a researcher throughout the entire process. I would be remiss not to thank my friends and family for their support and love.

Finally, I would like to recognize The International Research Foundation for English Language Education (TIRF) (tirfonline.org) for honoring this dissertation research with a doctoral dissertation grant (DDG) and for recognizing my research with the 2017 Russell N. Campbell Award, which is given to the top-ranked DDG in each annual competition. The grant provided important funding for my dissertation research.

Çanakkale, 2018

Özgür ŞAHAN

Abstract

The Impact of Rater Experience and Essay Quality on Rater Behavior and Scoring

This dissertation aimed to investigate the impact of rater experience and essay quality on rater behavior and scoring. In doing so, the variability of essay scores assigned to high-quality and low-quality essays were examined quantitatively while raters' decision-making strategies were investigated qualitatively. Using convergent parallel design as a mixed-methods approach, data were collected from 31 EFL instructors and two research assistants working at higher education institutions in Turkey. While 15 of the participants were from a specific university, the remaining participants represented various universities across Turkey. Based on their reported rating experience, participants were divided into three groups: low-experienced (*n* = 13), medium-experienced (*n* = 10), and high-experienced raters (*n* = 10).

Using an analytic scoring rubric, each participant assessed a number of 50 essays of two distinct qualities (high- and low-quality) and simultaneously recorded think-aloud protocols to determine the raters' decision-making processes while scoring EFL essays. In addition, raters' written explanations for their ratings were used to triangulate the verbal protocols. A total of 9,900 scores (1,650 total scores and 8,250 sub-scores), 446 think-aloud protocols, and 5,425 written score explanations were obtained from the participants. The analysis of quantitative data relied on generalizability (G-) theory approach as well as descriptive and inferential statistics; qualitative data were analyzed through deductive and inductive coding.

The results showed that high-experienced raters are more positive toward students' essays and assign higher scores compared to their less experienced peers. Furthermore, the high-experienced and low-experienced groups differed significantly in their total scores and mechanics component sub-scores assigned to low-quality essays. Additionally, G-theory analyses were conducted to determine the sources of measurement error and their relative contributions to the score variability. The results yielded a smaller rater effect when high- and low-quality essays were considered collectively, but it was found that raters contributed more to score variation when separate analyses were conducted for each essay quality. The qualitative findings suggested that raters in different experience groups display different decision-making behaviors while assessing essays of different proficiency levels. Overall, the findings provide striking insights for rater reliability in EFL writing assessment. Implications are discussed with respect to EFL writing assessment in the local and wider context from the perspective of fairness and rater reliability.

Keywords: EFL writing assessment, essay quality, generalizability theory, rater behavior, rater experience, score variability, think-aloud protocols

Özet

Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi

Bu tezin amacı, puanlayıcıların geçmiş puanlama tecrübeleri ve değerlendirilen kompozisyonların kalitesinin, puanlayıcıların değerlendirme esnasında sergiledikleri davranışlar ve kompozisyon puanları üzerindeki etkilerini araştırmaktır. İyi ve kötü seviyede yazılmış kompozisyonlara verilen puanlar, nicel araştırma yöntemine tabi tutulurken, puanlayıcıların karar verme stratejileri nitel olarak incelenmiştir. Araştırmada karma araştırma yöntemi olarak yakınsayan paralel karma yöntem deseni kullanılmıştır. Araştırma verisi, Türkiye'nin çeşitli üniversitelerinde çalışan 31 İngilizce okutmanı ve iki araştırma görevlisinden toplanmıştır. Katılımcıların 15'i aynı üniversitede görev yaparken, diğer katılımcıların görev yaptığı üniversiteler çeşitlilik göstermektedir. Araştırmada yer alan katılımcılar, geçmiş puanlama tecrübelerine bağlı olarak üç gruba ayrılmıştır. Buna göre düşük tecrübe grubu 13, orta tecrübe ve yüksek tecrübe grupları da 10'ar kişiden oluşmaktadır.

Her bir katılımcı, analitik puanlama ölçeği kullanarak iki farklı kalite grubundan oluşan 50 adet kompozisyon puanlamıştır. Bununla birlikte, sesli düşünme yöntemi kullanılarak katılımcıların İngilizce kompozisyon puanlarken sergiledikleri karar verme stratejileri incelenmiştir. Ayrıca, her bir puanlayıcı tarafından, verilen puanların gerekçelerinin belirtildiği yazılı açıklamalar sunulmuştur. Toplamda 9,900 adet kompozisyon puanı (1,650 toplam puan ve 8,250 alt puan), 466 adet sesli düşünme protokolü ve 5,425 adet yazılı değerlendirme puanı gerekçeleri elde edilmiştir. Nicel veriler, genellenebilirlik kuramı analizi ile birlikte, betimsel ve çıkarımsal istatistik kullanılarak analiz edilirken, nitel verilerin analizleri için tümdengelim ve tümevarım yöntemleriyle kodlama ve sınıflandırma yöntemi kullanılmıştır.

Araştırma bulguları, yüksek tecrübe grubunda yer alan puanlayıcıların öğrenci kompozisyonlarına karşı daha olumlu tutum sergilediklerini ve daha az tecrübeli puanlayıcılara göre daha yüksek not verdiklerini göstermektedir. Ayrıca, düşük kalitedeki kompozisyonlara verilen toplam puanlar ve buna ek olarak mekanik bileşenine (imla, noktalama ve büyük harf kullanımı) verdikleri puanlar göz önüne alındığında, yüksek tecrübe ve düşük tecrübe gruplarında yer alan puanlayıcıların birbirlerinden anlamlı bir şekilde farklılaştığı tespit edilmiştir. Ölçmedeki hata kaynaklarını belirlemek ve bunların, puan değişkenliğini ne ölçüde etkilediğini tespit etmek adına genellenebilirlik kuramı analizleri yapılmıştır. Analiz sonuçları, her iki kalitedeki kompozisyonlar bir arada düşünüldüğünde puanlayıcıdan kaynaklanan hatanın küçük olduğunu; ancak farklı kalitedeki kompozisyonlara verilen puanlar birbirinden bağımsız düşünüldüğünde, puanlayıcının puan değişkenliğine daha fazla katkıda bulunduğunu ortaya çıkarmıştır. Araştırmanın nitel bulguları dikkate alındığında ise, farklı tecrübe düzeylerine sahip puanlayıcıların yüksek ve düşük kalitedeki kompozisyonları değerlendirirken farklı karar verme stratejileri uyguladıkları belirlenmiştir. Bu tez araştırması genel olarak, İngilizce yazma becerisinin değerlendirilmesi alanındaki puanlayıcı güvenirliği konusunda çarpıcı sonuçlar ortaya koymaktadır. Kurumsal ve daha genel bağlamlar düzeyinde araştırmanın bulgularının etkileri tartışılmaktadır.

Anahtar Kelimeler: genellenebilirlik kuramı, İngilizce kompozisyon değerlendirme, kompozisyon kalitesi, puan değişkenliği, puanlayıcı davranışı, puanlayıcı deneyimi, sesli düşünme protokolü

Table of Contents

List of Tables

List of Figures

Figure **Title** **P**

Chapter I

Introduction

It is a well-known fact that foreign language skills are important in an increasingly globalized world in that language tests play a crucial role in people's lives, opening doors to opportunities in education, business, and even moving to other countries (McNamara, 2000). In other words, when the acquisition of a language skill becomes important, testing that skill gains significance as well (Weigle, 2002) in order to track the continual development of learners and make high-stakes decisions relying on assessment outcomes. In this sense, it is of great importance to administer effective testing methods that allow individuals to perform at the required standard of language use (Fulcher, 2010).

Assessment can simply be defined as different ways used to gather information on learners' language abilities (Hyland, 2003). As for writing, an assessment task is the process in which students generate a piece of writing, and it is known to be "the most common method for writing assessment in both first- and second-language contexts" (Weigle, 2002, p. 58). In writing performance tests, students are expected to produce a satisfactory amount of writing and experienced raters make judgements about the product relying on agreed-upon criteria, which represent the quality of their performance (McNamara, 2000). In this regard, if the interpretation of a score assigned to a test is an indicator of an individual's performance, that score should be reliable and valid (Bachman, 1990). However, assessing students' English as a second language (ESL) or English as a foreign language (EFL) writing performance is not only assigning scores to their essays, but rather a complex process that brings several factors together to compose a decision about students' performance (Weigle, 2002). To put it differently, a score assigned to the essay is not the outcome of the interaction that occurs between test-taker and the test, but the result of the interactions among several factors including the test-taker, the prompt or task, the written text itself, the rater(s), and the rating scale (Hamp-Lyons, 1990; Kenyon, 1992; McNamara, 1996). Therefore, ratings given to individuals' performances are believed to be subjective since they are not only reflections of the quality of performances but also the quality of raters' judgements (McNamara, 2000), which puts the rater and rating process in a central place (Attali, 2015).

Given the multiple aforementioned factors contributing to variability in essay scores, assessing writing is a complicated and challenging process (Barkaoui, 2008; Fulcher, 2010; Hamp-Lyons, 1991; Huang, 2007, 2008, 2009, 2011; Huang & Foote, 2010; Huang & Han, 2013; Hughes, 2003; Weir, 2005). Therefore, it is not very likely to obtain perfect score reliability in writing assessments (Bachman, 1990; Hughes, 2003). Demanding high reliability is, however, natural especially when important decisions about the learners are made from the writing scores. As such, several factors related to test design, test administration, and scoring should be treated appropriately in order to make the tests more reliable (Hughes, 2003).

In this sense, rater training and previous experience of the raters are considered as effective factors to ensure reliability of scores in terms of the aspects of intra- and inter-rater consistency. Therefore, investigating raters' scoring background can help reduce the variability in essay scores (Carlson, Bridgeman, Camp, & Waanders, 1985; Cumming, 1990; Hamp-Lyons, 1990; Homburg, 1984; Myers, 1980; Najimy, 1981; Reid, 1993; Upshur & Turner, 1995).

Problem Statement

Among other performance assessments, assessing EFL writing performance is frequently carried out for three main purposes in Turkish universities: 1) to address the high-stakes entrance and exit exams given to students in the one-year, intensive English preparatory programs that universities generally offer to students from different majors prior to the start of their departmental courses and which are required for students enrolled in English-medium departments; 2) to assess students' writing performances throughout their university education in order to evaluate their progress in English; and 3) to evaluate students' writing performance as a prerequisite for exchange programs like Erasmus+, since such programs require a good command of EFL writing. Given that each of these purposes influences students' progress in higher education, it is essential to assess students' written production in a reliable way.

The testing offices of the English language preparatory programs at universities are mostly responsible for preparing exams for the aforementioned purposes. However, scoring procedures do not always follow formal, predetermined steps, such as training and calibrating raters to rate the essays reliably. As such, different assessment protocols are implemented at different institutions. To illustrate, in some cases only a single rater assigns scores to students' written products based on his/her impressions and inner-criteria while in other cases double-grading is employed with a reliable rubric. Furthermore, anonymous evaluations are adopted at some institutions while the transparency of students' identities may manipulate the assessment processes at other institutions. The application of different scoring preferences and procedures at different institutions or within the same institution contributes to unfair judgement and unfortunately is the norm across Turkish universities. Therefore, there is a need for a standardized and sound assessment system in order to provide students with fair scorings.

The aforementioned discussions about reliability issues in assessing writing are drawn from multiple factors that contribute to the fairness of writing scores (Gebril, 2009; Han, 2013; Huang, 2011; Huang & Foote, 2010, Saeidi & Rashvand Semiyari, 2011). Generally, the variability of scores stems from three sources: a) students, b) rater types, and c) writing tasks (Barkaoui, 2007a; Elorbany & Huang, 2012; Gebril, 2009, 2010). However, many other factors have been identified in the literature as affecting the writing scores assigned by raters to a single task, including rating mode, scoring method, and rater training (Barkaoui, 2008; Brown, 1991; Cumming, 1990; Huang, 2011; Lumley, 2005; Weigle, 2002). Of these factors, rater variation is considered most central to writing performance assessment (Huang, 2011; Huang & Foote, 2010; Huot, 1990; Lim, 2011; Wolfe, 2005). Raters show a variety of differences in terms of their professional experience, linguistic background, educational background, expectations and beliefs, and tolerance for error (Weigle, 2002). These variations cause the assignment of differing scores to the same essays by different raters or the fluctuation in scores to the same essays by the same raters at different times (Homburg, 1984; Huang, 2011; Huot, 1990).

As one of the rater features, previous rating experience is attributed to ensuring fair judgment, placing expert scorers in a superior position throughout the evaluation processes. Yet, expertise in assessing writing does not necessarily promise reliable scores. Additionally, the contrast effect in rating while assessing papers of different qualities simultaneously is worthy of discussion in that while a medium quality essay tends to receive a low score when it is assessed after reading several high-quality essays, it tends to receive a higher score when it is preceded by a number of lower quality essays (Daly & Dickson-Markman, 1982; Freedman, 1981; Hughes & Keeling, 1984). Therefore, it is essential to understand the differences and commonalities in raters' reactions to essays of different qualities in order to better understand the variability of ratings. To this end, this research study focuses on two factors, namely scorers' rating experience and essay quality, to investigate their impact on the variability of EFL essay scores and rating behaviors that the raters exhibit in Turkish tertiary-level education. Given that assessment problems have been under-researched at the institutional and national levels in Turkey, this research gains significance by investigating two main sources of error in EFL writing assessment to establish meaningful and generalizable measurements that should be relevant beyond individual contexts.

Purpose of the Study

The main purpose of this study was to investigate the impact of rater experience and essay quality on rater behavior and essay scores. First, it aimed to explore whether professional experience matters for the variability and reliability of ratings by examining raters with varying rating experience. Second, this study attempted to observe the different behaviors and decision-making strategies that raters exhibit while assessing essays of different qualities. Third, this study aimed to measure the extent to which the aforementioned factors contribute to the variability and reliability of EFL writing scores. Adopting a mixed-methods research design, the variability and reliability of ratings assigned to the essays were examined quantitatively through the employment of G-theory approach. Qualitative data were collected through think-aloud protocols and written score explanations to investigate the decision-making behaviors of the raters.

From the quantitative perspective, the first set of questions were as follows:

1. Are there any significant differences among the analytic scores of the low- and high-quality EFL essays?

2. Are there any significant differences among the analytic scores assigned by raters with varying previous rating experience?

3. What are the sources of score variation that contribute most (relatively) to the score variability of the analytic scores of EFL essays?

4. Does the reliability (e.g., dependability coefficients for criterion-referenced score interpretations and generalizability coefficients for norm-referenced scores interpretations) of the analytic scores of raters differ based on their amount of experience?

Moreover, based on the qualitative data, the following questions were asked:

5. How do raters make decisions while rating different quality EFL essays analytically?
6. How is rating experience related to EFL raters' decision-making processes and the aspects of writing they attend to?

Significance of the Study

The rater is located at the heart of the assessment process (Lumley, 2005) and one of the rater factors that seems to play a prominent role in the assessment process is the raters' professional experience (Barkaoui, 2010a). Although the findings of previous studies are contradictory, empirical research has investigated the effect of professional experience on essay scores (Barkaoui, 2008, 2010a, 2010c, 2011a; Cumming, 1990; Hamp-Lyons, 1996; Leckie & Baird, 2011; Lim, 2011; Lumley & McNamara, 1995; Reid & O'Brien, 1981; Rinnert & Kobayashi, 2001; Shohamy, Gordon, & Kraemer, 1992; Song & Caruso, 1996; Sweedler-Brown, 1985; Weigle, 1999; Wolfe, 2005). In some cases, there was a positive correlation between rater experience and rater leniency (e.g. Song & Caruso, 1996) while the reverse was reveled in other studies (e.g. Barkaoui, 2010a). Further, sometimes no significant difference was found between less experienced and more experienced rater groups (e.g. Shohamy et al., 1992) in terms of reliability, while experienced raters were found to be more reliable in some cases (e.g. Reid & O'Brien, 1981). Added to that, experienced raters and novice raters might show differences in employing different decision-making strategies (e.g. Cumming, 1990). Since previous research has mostly examined the differences between novice and experienced raters and revealed conflicting results, investigating the impact of varying previous rating experience of raters in Turkish context gains significance.

Another inspiration of the current study is the limited amount of research that has been carried out to examine the impact of essay quality on rating variability and reliability of ESL/EFL writing. Learners' expertise in the second language (L2) and their first language (L1) backgrounds have been the main concerns of some studies (Baba, 2009; Brown, 1991; Han, 2017; Huang, 2008; Huang, Han, Tavano, & Hairston, 2014; Song & Caruso, 1996). For example, Brown (1991) found no significant difference in the scores assigned to the essays written by native English speaker (NES) and EFL students. However, Huang (2008) revealed that ESL students had lower scores than NES students given their linguistic deficiencies. When it comes to writer proficiency, raters tended to give more consistent scores to high-quality essays (Han, 2017; Huang et al., 2014). The research examining scoring differences between papers of distinct qualities is limited and mostly restricted to ESL contents. Therefore, this study is significant given that it attempts to fill the research gap related to the impact of essay quality on writing performance assessment.

Another important consideration of this research is related to its methodology. Most quantitative research relies on classical test theory, which is considered a weak theory as it accounts for only a single source of variance within a given analysis (Huang, 2008, 2012; Linn & Burton, 1994). However, this study uses G-theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) as a theoretical framework for quantitative analysis to detect rater variation and reliability of writing assessment because of its sophisticated nature to detect multiple sources of variability on essay scores (Shavelson & Webb, 1991). Additionally, this study employs think-aloud protocols (TAPs) to collect qualitative data for comparing the evaluation criteria and the rating processes of raters in different experience groups.

As a result, the current study attempts to examine the rater variation in EFL writing assessment in Turkey by putting specific emphasis on previous experience in writing assessment and the role of essay quality in the variability of ratings. In order to understand what deviates raters from each other cognitively, this study strives to explore the differences among raters' decision-making processes even if they are expected to use the same scoring criteria. The findings will put forward suggestions and implications for the standardization of writing assessment situations and regulating institutional assessment protocols.

Definitions of Key Terms

The key terms that are considered central to the purpose of this study are listed as follows:

EFL students – students whose L1 is Turkish and learn English as a foreign language.

Writing assessment – it implies evaluating an essay by assigning a score to the written performance and commenting on it in the context of the study. A variety of terms including "grading," "marking," "rating" and "scoring" can be used interchangeably to refer to the assessment process.

Rater – it refers to the assessor grading ESL/EFL writing, implying EFL instructors working at higher education institutions in Turkey.

Previous rating experience – the number of years that a rater has spent rating EFL/ESL writing professionally.

Rater behavior – in the context of the study, rater behavior refers to different ways by which a rater arrives at a decision about students' written performance (Huot, 1990).

Holistic scoring – assessing a writing sample by assigning one score to reflect the overall quality of the paper (e.g., grammar, content, organization, style and quality of expressions, and mechanics).

Analytic scoring – the process of evaluating each component of writing performance such as grammar, content, organization, style and quality of expressions, and mechanics separately using a rating scale.

Object of measurement – it refers to the entity under investigation. In this context, the objects of measurement are students (Shavelson & Webb, 1991).

Facet – a particular aspect of a measurement procedure (Shavelson & Webb, 1991). In the context of this study, rater and essay quality can be defined as potential sources of measurement error.

Condition – levels of a facet (e.g., for the facet rater: rater 1, rater 2, etc.; Shavelson & Webb, 1991).

Universe – it represents the overall number of conditions of a facet or amalgamation of facets (as in an interaction) (e.g., universe of items, universe of raters, and universe of items-raters) (Shavelson & Webb, 1991).

Universe of generalization – it refers to the universe of conditions of a facet to which a decision-maker wants to generalize (Shavelson & Webb, 1991).

Universe score – the value attributed to a person's observed scores over all observations in the universe of generalization. It is also known as "true score" in classical test theory (Shavelson & Webb, 1991).

Generalizability- (G-) study – a type of study in G-theory to evaluate the relative importance of various sources of measurement error and investigate the effects of diverse changes in the measurement design (e.g., different number of tasks or raters/ratings; Brennan, 2001b).

Decision- (D-) study – a type of study which integrates the ideal design to allow the interpretation of score reliability in the norm-referenced or criterion-referenced frame of reference (Brennan, 2001b).

Variance component – refers to the facet(s) that has an effect size in a G-study. It helps the investigator estimate the magnitude of explained variance components within the given design. It not only accounts for each variance component but also explains the percentage of variance resulting from the interactions between facets (Shavelson & Webb, 1991).

Fixed facet versus random facet – If the researcher is dealing with the instances under investigation and does not desire to generalize beyond those instances, then the facet is treated as fixed while all conditions in a facet are exchangeable with the ones in the universe when the facet is considered random (Briesch, Swaminathan, Welsh, & Chafouleas, 2014; Güler, Uyanık, & Teker, 2012).

Norm-referenced versus criterion-referenced – the scores of each test-taker are interpreted relative to the other test-takers' performance in a norm-referenced test. In a criterion-referenced test context, each test-taker's score is interpreted relative to a fixed set of predetermined test criteria (Brown, 1996).

Relative error versus absolute – There are two types of decisions made in measurement theory: relative and absolute. While the former deals with an individual's relative position in a population, the latter concerns the individual's level of knowledge, skills, and attitudes regardless of others' performance. In this sense, the error related to relative decision is considered relative error, and absolute error is the result of error associated with absolute decision (Shavelson & Webb, 1991).

Generalizability coefficient versus dependability coefficient – generalizability coefficients are used in a norm-referenced score interpretation and are denoted by Ep^2 or *G*- whereas dependability coefficients are used in a criterion-referenced score interpretation and are denoted by Φ (Briesch et al., 2014; Shavelson & Webb, 1991).

A generalizability coefficient is the ratio of the universe score variance to itself plus relative error variance ($Ep^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2}$). It is the analogue of a reliability coefficient in classical theory. A dependability coefficient is the ratio of the universe score variance to itself plus absolute error variance ($\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2}$) (Brennan, 2001a, p.13).

Organization of the Dissertation

This dissertation consists of five chapters including introduction, literature review, methodology, results, and conclusion and discussion. The following chapter gives a detailed review of the relevant literature in a deductive manner beginning with general information about L2 writing assessment and elaborates into the details and the factors effecting writing performance assessments. After scrutinizing the relevant studies about raters' professional experience and essay quality as the factors influencing writing score variability and reliability, the section continues with a summary of rater cognition in terms of decision-making behaviors. Thereafter, Chapter 2 ends with a summary and statement of research gaps in EFL writing assessment. Chapter 3 starts with the research design and theoretical framework. Following that, descriptions of participants, data collection instruments, data collection process, and data preparation procedures for the analysis are presented respectively. In Chapter 4, the results are organized and reported to answer the qualitative and quantitative research questions separately. Finally, Chapter 5 summarizes and discusses the findings, explains the limitations of the study and touches upon pedagogical and methodological implications in the end.

Chapter II

Literature Review

This chapter reviews the research literature regarding the factors that impact the variability and reliability of ESL/EFL writing performance assessment. First, brief information is given about L2 writing assessment and the section continues with an overview of EFL writing assessment in higher education in Turkey. Second, reliability, validity, and fairness issues concerning writing performance evaluation and the factors effecting the reliability and variability of essay scores are discussed through the review of empirical studies. Third, the impact of rater's professional experience and essay quality on ESL/EFL writing scores are examined respectively. Finally, rater cognition in terms of EFL/ESL writing assessment is discussed followed by a summary and statement of research gaps in EFL writing assessment and the significance of this dissertation to bridge those gaps.

Second Language Writing Assessment

Performance tests are the most commonly used tests in writing in which individuals are expected to produce a satisfactory piece of writing upon which experienced raters make judgements relying on an agreed-upon criteria (McNamara, 2000; Saeidi & Rashvand Semiyari, 2011). This assessment type is employed to design a testing procedure that includes the observation or simulation of real-world behavior and activity from which the raters evaluate the performance (Weigle, 2002), providing the advantage of direct assessment of learners' productive language skills (Johnson & Lim, 2009). McNamara (1996) makes a distinction between *strong sense* and *weak sense* of performance assessment in language testing. While fulfilment of the task is prioritized in strong sense of performance assessments, the focus is on the language use in the weak sense. McNamara asserts that most language tests appear somewhere in between these two extremes of the continuum.

Regardless of where performance assessment stands in the language testing process, the primary purpose of testing in an educational setting is to provide information for decision-making processes (Bachman, 1990). In this regard, assessment purposes can generally be described in five categories including a) *placement:* allocating students to appropriate language proficiency groups; b) *diagnostic*: identifying students' needs based on their strengths and weaknesses; c) *achievement*: exploring students' progress considering the course objectives and outcomes; d) *performance*: finding out students' success in performing specific tasks; and e) *proficiency*: measuring students' general language proficiency levels (Hughes, 2003; Hyland, 2003). Specifically, assessing writing has two main purposes: 1) making inferences out of test performance and 2) making high-stakes and low-stakes decisions based on those inferences (Bachman & Palmer, 2010; Hughes, 2003; Weigle, 2002). To illustrate the difference between high- and low-stakes decisions, placing students in a level group based on performance assessments can be considered to have a minor impact on students' lives (low-stakes) while awarding university admission or a scholarship based on performance assessment may impact the students' futures significantly (high-stakes) (Weigle, 2002). Therefore, reliability and fairness should be ensured in writing performance assessments given the notable impact that the outcome of the rating process has on individuals' lives (Attali, 2015; Baker, 2010).

In parallel with the assessment purposes mentioned above, another distinction can be made between *formative* and *summative* assessment in that both assessment types carry different purposes in the testing process. Formative assessment aims to verify to what extent learners have progressed in achieving the learning objectives and outcomes and provides information to modify future learning situations, while summative assessment intends to measure the ultimate success of the students at the end of the process (Fulcher, 2010; Hughes, 2003). Referring to the discussion on the tests being high-stakes or low-stakes in the previous paragraph, summative assessment results can be considered respectively more important in terms of the generalizable meaning of scores in that high-stakes tests are used to certify an ability and compare the performance of different educational settings across the world (Hughes, 2003).

As for the characteristics of a writing assessment task, Hamp-Lyons (1991) suggests a list of items involving a minimum-sized of a piece of writing (100 words suggested), writing prompt, scoring scale, rater and grade. Weigle (2002) extends this list by adding two more items: 1) a limited time frame and 2) confidentiality of the topic which should be unknown to the test-takers before the test. Although all the aforementioned facets of an assessment task are included in the evaluation process, the performance of students tends to vary from each other due to a variety of factors. The factors contributing to score variances can be attributed to two general sources: a) meaningful variance created by the purposes of a test and b) measurement error or error variance generated by extraneous sources (Brown, 2005). These sources of variance can be listed as environmental factors (i.e. noise, lighting, weather, space, and location), test administration processes (i.e. direction, equipment, and timing), test-taker variables (i.e. health, fatigue, motivation, concentration, and testwiseness), scoring procedures (i.e. errors in scoring, subjectivity, and rater biases), and the test and test items (i.e. item types, number of items, test booklet clarity, answer sheet format, and particular sample of items), all of which impact the reliability of scorings assigned to students’ test performances (Brown, 2005; Schoonen, 2005).

Considering the importance of accountability in assessment, fair judgement plays a chief role in establishing meaningful and generalizable measurements that should be relevant beyond individual contexts. As such, the following section discusses reliability, validity, and fairness concerns specifically related to L2 writing assessment and provides an overview of the

EFL writing assessment situations at Turkish universities followed by a review of factors affecting the reliability and validity of ratings in EFL/ESL writing.

Reliability, validity, and fairness. The discrimination between subjective and objective tests can be entirely made in terms of scoring procedures (Bachman, 1990). While the correctness of a test taker’s responses to the questions is subject to a pre-determined answer key in an objective test, the interpretation of the scorers in a subjective test determines whether the answer is correct or not (Bachman, 1990). That the evaluation of students’ performances is dependent on the raters’ interpretations in subjective tests raises concerns about the reliability of performance assessments. This is particularly true in writing assessments, which can be considered subjective in nature. In other words, a score assigned to the essay is not the outcome of the interaction that occurs between test-taker and the test, but the result of the interactions among several factors including the test-taker, the prompt or task, the written text itself, the rater(s), and the rating scale (Hamp-Lyons, 1990; Kenyon, 1992; McNamara, 1996), which can lead to erroneous measurements in writing assessment. Therefore, reliability investigates the extent to which a test score—an individual’s performance—is affected by measurement error rather than the language ability that is intended to be measured (Bachman, 1990). That is to say, reliability examines how to derive consistent scores across different raters, times, settings, test forms, and other characteristics of measurement (Bachman, 1990; Weigle, 2002).

There are two main types of inconsistencies in scoring: a) inter-rater reliability is concerned with the consistency of scores assigned to same essays by two raters; b) intra-rater reliability is related to the inconsistencies in the scores of a single rater assigned to the same essay or essays in similar quality across different times (H. D. Brown, 2004; J. D. Brown, 2005; Shohamy et al., 1992; Weigle, 2002). Although it is difficult to ensure inter-rater reliability because of rater variables—hence full inter-rater consistency does not seem to be promised no matter how the raters are trained on scoring essays—Weigle (2002) suggests several procedures that can be implemented in order to improve inter-rater reliability in writing assessment as follows:

> [D]esigning and pre-testing prompts carefully to make them accessible to all test-takers, selecting and training raters, double-marking of essays, ensuring the independence of scores so that one rater is not influenced by the scores that another rater gives, and using a scoring rubric along with model essays that instantiate the criteria outlined in the rubric. (Weigle, 2002, p. 59)

In addition to the procedures underlined above, White (1994) proposes three more practices to maintain reliability in writing assessment: a) scoring essays in a controlled reading which means that raters come together in the same place and time to grade essays; b) checking on the scoring in progress to make sure the individual raters are following the pre-determined scoring standards; and c) evaluating the scoring process and keeping records pertaining to the assessment tasks in order to discriminate between reliable and unreliable raters for the following grading sessions. Absolute reliability is not likely to be achievable (Hughes, 2003; Hyland, 2003) even if the required measures are taken in order to form a safe assessment context. However, greater reliability should be demanded when important decisions are to be made from test scores (Hughes, 2003).

Reliability is an essential consideration in testing and an important requirement for test validity (Bachman, 1990; Weigle, 2002). Reliability is concerned with the quality of test scores while validity is related to the quality of test interpretations and purpose (Bachman, 1990). That is to say, validity examines whether the inferences and decisions made from test scores are meaningful, appropriate and useful regarding the purpose of the test in that there must be a high level of certainty that a test score is the indicator of particular individual's ability (Bachman, 1990). Instead of approaching reliability and validity as different constructs, they should be considered to be complementary because both of their concerns are essentially to "minimize the effects of measurement error" and "maximize the effects of the language abilities we want to measure" (Bachman, 1990, p.161).

Statistically speaking, the reliability coefficient varies between 0 and 1, allowing test designers to assess the reliability of tests and scorings in that ideal reliability is close to its maximum (1) when consistent results are obtained for a particular group of examinees (Bachman, 1990; Fulcher & Davidson, 2007; Hughes, 2003; Kane, 2008). With respect to the reliability coefficient of tests, it is inevitable to touch upon the distinction between unobservable behaviors—language-abilities in our context—and observed test scores; in other words, given that the language abilities being measured are abstract, they are not subject to direct observations which can yield an individual's 'true' score for a given ability (Bachman, 1990, p. 166).

An observed score is comprised of two components: "a true score" which represents an individual's level of ability and "an error score" that stems from the factors other than the ability being tested (Huot, 1990; Bachman, 1990, p. 167; Fulcher & Davidson, 2007, p. 104). According to *classical true measurement theory*, the variance of the observed scores is equal to the sum of variances of true scores and error scores, the former of which rely on the difference between performances of examinees while the latter refers to the unsystematic and random measurement error (Bachman, 1990; Huang, 2009; Hughes, 2003; Huot, 1990). In this sense, the greater the proportion of true score and the less the contribution of error score are, the more reliable the observed score attained can be considered.

In a nutshell, fairness is desired in every assessment situation; however, fairness is hard to achieve due to the contribution of several factors to the reliability of writing scores (Bachman & Palmer, 2010; Breland, 1983; Han, 2013; Huang, 2007, 2008, 2009, 2011; Huang & Foote, 2010). As such, several precautions including training the raters, double-grading the essays, using scoring rubrics etc., have been put forward to increase the reliability of ESL/EFL writing assessment (Bachman & Palmer, 2010; Hughes, 2003; Weigle, 2002; White, 1994). Therefore, previous research has investigated the impact of the aforementioned factors on the reliability of ESL/EFL writing assessment and the effectiveness of various measures to eliminate reliability concerns in the assessment process (Attali, 2015; Bacha, 2001; Barkaoui, 2007b; 2008; 2010a, 2010b, 2010c; Ebel & Frisbie, 1991; Elorbany & Huang, 2012; Han, 2013; Henning, 1991; Huang, 2008, 2011; Huang & Foote, 2010; Shi, 2001; Song & Caruso, 1996; Sweedler-Brown, 1985; Weigle, 2002).

An overview of EFL writing assessment in tertiary education in Turkey. University admissions for undergraduate study in Turkey rely on a national-level examination administered by the Student Selection and Placement Center (ÖSYM, abbreviated in Turkish). The only way to enter a university for prospective university students is to take the aforementioned exam, and test-takers are ranked based on their exam scores. The candidate students are placed in a university from their list of preferences according to their scores and rankings.

The medium of instruction at tertiary education in Turkey is 100% Turkish, 30% English or 100% English, and is determined by the English language proficiency of individual departments' academic staff. According to the regulations of Turkish Council of Higher Education, university students have to meet English language requirements in order to start their departmental studies in programs that are partly (30%) or completely (100%) in English.

Every university administers its own English proficiency tests that are generated by the testing unit of the university's English preparatory program or provided by international English education and publishing companies at the beginning of the first academic year. However, these high-stakes tests differ from each other in some aspects including question type, question difficulty, and test content in terms of skills assessed based on the specific goals and educational policies of each university. If students meet the predetermined performance standards of these tests, they can start their education in their departments. Otherwise, students have to receive foundation English courses to develop their reading, writing, listening, and speaking skills at an English preparatory program, in which students are placed in distinct language proficiency levels decided by a placement test. In English language preparatory programs in Turkey, writing constitutes a fundamental portion of the English language curriculum because the ability to write effectively has become more valuable, especially in higher education, given that writing is considered not only as a standardized way of communication but also as a key to successful learning (Weigle, 2002). Therefore, students receive writing classes in each level of their English preparatory year during which they learn to write beginning with sentence-level texts and advancing to paragraphs and essays.

In addition to existing as a separate course in the preparatory programs, writing has become a common way to test students' English language achievement in different courses with activities such as paraphrasing in reading courses or note-taking in listening courses. Several courses aiming to teach writing skills, such as Advanced Reading and Writing Skills and Academic Writing, are also offered in departments such as English Language Teaching (ELT), English Language and Literature (ELL), English Linguistics (EL), and Translation Studies (TS). Additionally, teaching how to write in English is an interdisciplinary concern both at undergraduate and graduate schools of Turkish universities due to the status of English as an international language in science and academy.

Considering the crucial role of writing skill instruction as stated above, assessing that skill is equally as important. The best way to test learners' writing ability is to "get them to write" (Hughes, 2003, p. 83). However, a predetermined and standardized way is not followed while assessing learners' writing performance, potentially causing score variations and unfair judgements. Instead, various assessment protocols are implemented at different universities across the country. While some universities rely on a double-grading system using a scoring rubric, only a single rater judges students' essays at other universities. On the one hand, essays are graded anonymously at some institutions, but on the other hand, raters' at other institutions know students' identities, which potentially biases scoring. These diverse implications give rise to assessment reliability issues not only among different universities but also within the same university. Therefore, it is indisputable that there are a number of reliability concerns stemming from the assessment procedures of EFL writing performances, which are used to make high-stakes decisions that impact students' lives significantly.

Factors Affecting ESL/EFL Writing Assessment

Given that test performance is affected by factors other than the skills being tested, identifying the potential sources of error in a measurement is a fundamental concern in language test development and use (Bachman, 1990). Weigle (2002) benefits from the works of McNamara (1996) and Kenyon (1992) to give a list of the factors that have an impact on test scores as follows: task variables, text variables, rater variables, rating scales, context variables, and test-taker variables. Among these variables, rater variation is considered central to writing performance assessment (Bachman, 1990; Cooper, 1984; Huang, 2008, 2011; Huang & Foote, 2010; Huot, 1990; Lim, 2011; Stalnaker & Stalnaker, 1934) as raters bring their experience, expectations, background, and values to the assessment process (Huang, 2009; Weigle, 2002). Rater variations contribute to the assignment of varying scores to the same essays by different raters or fluctuation in score to the same essays by the same raters at different times (Bachman, 1990; Homburg, 1984; Huang, 2008, 2009, 2011; Huot, 1990). Rater variables have been investigated based on two main foci, which are the components of essays that raters attend to while assessing writing and the impact of rater characteristics on essay scores (Weigle, 2002).

An extensive body of studies on the factors including writing task and essay topic, rater's previous scoring experience, rater's L1, rater training, rating methods, and sociocultural aspect and purpose of assessment is analyzed in the forthcoming sections of this chapter. Additionally, rater's rating experience and essay quality are elaborated on as the two factors researched in this dissertation. Furthermore, benefiting from the findings of a number of empirical studies, this section explains rater cognition and the decision-making strategies that are applied during writing skill evaluations. Table 1 shows a brief summary of reviewed studies on the aforementioned factors and issues involved in L2 writing assessment. Some of the studies were presented under two or more categories because of their multiple foci relevant to this research.

Table 1

Summary of Reviewed Empirical Studies

Focus of the study	**Number of studies**	**Studies reviewed**
Writing Task and Essay Topic	5	Gebril (2009); Hamp-Lyons & Mathias (1994); Jennings et al. (1999); Saeidi & Rashvand Semiyari (2011); Weigle (1999)
Rater's Professional Background	3	Elorbany & Huang (2012); Santos (1988); Song & Caruso (1996)
Rater's L1	3	Johnson & Lim (2009); Kobayashi (1992); Shi (2001)
Rating Methods	6	Bacha (2001); Barkaoui (2007b); Barkaoui (2010c); Han (2013); Knoch (2009); Song & Caruso (1996)
Rater Training	5	Attali (2015); Knoch et al. (2007); Shohamy et al. (1992); Sweedler-Brown (1985); Weigle (1994)
Socio-political Aspect	1	Baker (2010)
Rater's Rating Experience	8	Barkaoui (2010a); Cumming (1990); Leckie & Baird (2011); Lim (2011); Rinnert & Kobayashi (2001); Song & Caruso (1996); Wolfe (2005); Wolfe, Kao, & Ranney (1998)
Essay Quality	6	Brown (1991); Engber (1995); Ferris (1994); Han (2017); Huang (2008); Huang et al. (2014)
Rater Cognition	13	Baker (2012); Barkaoui (2007b; 2010c) Cumming et al. (2002); DeRemer (1998); Eckes (2008); Freddman & Calfee (1983); Frederiksen (1992); Han (2017); Lumley (2002); Vaughan (1991); Wolfe (2005); Wolfe & Feltovich (1994)
Total	**50**	

Writing task and essay topic. It should be noted that issues related to the writing topic such as test-takers' interest in the topic, their prior knowledge of the topic or whether the topic overlaps with the opinions of the test-takers may have an impact on examinees' writing test performance (Jennings, Fox, Graves, & Shohamy, 1999). Added to that, the number of tasks administered to the students at once has been debated and it has been contended that generally only one or two tasks can be administered to the students at the same time because of the practical considerations such as test administration time and cost of scoring (Weigle, 1999). Therefore, the tasks should be designed in a way that allows all of the test-takers to show their performance equally at the maximum level (Hamp-Lyons & Mathias, 1994; Weigle, 1999, 2002). When the examinees are given respectively lower scores for their written performance, it is generally attributed to the difficulty of that specific writing task or prompt (Weigle, 1999). However, the relationship between task difficulty and scores is not simple to explain given that rater-essay prompt interaction may affect the scoring process (Weigle, 1999). In this sense, several studies focused on the impact of writing task and essay topic on the variability of essay scores from various perspectives including the difficulty of the prompt (Hamp-Lyons & Mathias; 1994), rater-prompt interaction (Weigle, 1999), the impact of students' choice or no-choice of topic on their performance (Jennings et al., 1999), and comparison of different task types (Gebril, 2009; Han, 2013; Saeidi & Rashvand Semiyari, 2011).

Hamp-Lyons and Mathias (1994) examined the extent to which experienced raters' judgments of prompt difficulty varied. Four expert raters were asked to rate the difficulty of 64 prompts that were used for the writing section of the Michigan English Language Assessment Battery (MELAB). The results showed that most of the essay prompts were considered to be of moderate difficulty when the ratings of the four scorers were summed. Additionally, the study looked at the relationship between prompt difficulty and mean writing scores by examining previous scores assigned to the prompts under investigation in 8,583 cases. In contrast to what had been expected, the findings revealed a positive correlation between the judged prompt difficulty and the mean writing scores. That is to say, as the difficulty of prompts increased, mean writing scores increased as well. In this study, researchers also developed five prompt categories including expository/private, expository/public, argumentative/private, argumentative/public, and a combination of two or more of the four types. Moreover, the relationship between prompt category and the mean writing scores was researched and the results showed a reverse correlation in that the easiest prompt category considered by the raters—expository/private—received the lowest mean scores while the most difficult which was argumentative/public prompt was assigned with the highest mean scores.

Using quantitative and qualitative approaches, Weigle (1999) investigated the rater-prompt interaction in the setting of English as a Second Language Placement Examination (ELSPE). The writing section of the test includes two prompts—graph reading and choice justification—and examinees write on one of these prompt in 50 minutes. Sixty essays from an earlier set of compositions that represent each topic evenly were chosen for this study. Relying on ELPSE rubric consisting of three subscales (content, rhetorical control, and language), experienced and inexperienced raters were asked to rate the essays while thinking-aloud. The results did not indicate any significant differences between the two groups of raters' scores assigned to choice essay; albeit, inexperienced raters graded the graph essays more harshly than the experienced raters did. Following a training session, the differences found between the two groups were eliminated though. The analysis of verbal protocols revealed the possible reasons triggering the observed differences between the groups as follows: 1) the descriptors of the rubric used in this study may have not addressed the characteristics of graph essay, leading inexperienced raters to make harsh judgements, and 2) the nature of other choice essay types elicits standard responses that match the generic essay structure that raters expect; however, examinees can produce answers to graph reading tasks which the scorers may not be familiar with, causing raters to refer back to their prior rating experiences.

In the same year, Jennings et al. (1999) investigated whether students performed differently when they were given a choice of topic than those who did not have any choice of topic within the context of Canadian Academic English Language (CAEL) Assessment. In doing so, 254 ESL students hailing from numerous backgrounds and applying to Canadian universities participated in this study, and they were randomly and evenly assigned to two conditions: choice of topic and no choice of topic. According to the results, there was no significant difference between the performances of the two groups of examinees. However, though not statistically significant, the mean scores for the group that chose its topic were slightly higher than those of no-choice group. Furthermore, both groups identified time as the most important factor impacting their performance and the topic was reported as the second most important factor.

Investigating task type rather than topic choice, Gebril (2009) examined the score generalizability of independent and integrated (reading-based) writing tasks. Three experienced NES raters were given a set of essays collected from 115 Egyptian university students studying English language teaching. The essays were written on four tasks evenly representing independent and integrated categories. The raters carried out the assessment process by using separate holistic rubrics for each category. The results demonstrated that students performed similarly on both tasks while essays on integrated tasks received slightly higher mean scores. On the other hand, the generalizability analysis identified the triple interaction of persons (examinees), raters, and tasks as the largest variance component, followed by persons and person-by-task effect in both task types. When the data were further analyzed, it was seen that the absolute error variance would be reduced most when the number of tasks and raters were increased from one to two. Finally, because integrated tasks produced scores as reliable as the scores derived from independent tasks, this study suggested that reading-to-write tasks could be used more commonly in writing assessment.

Considering the impact of task from a different perspective, Saeidi and Rashvand Semiyari (2011) researched the impact of rating methods and task types on EFL learners' writing scores. In doing so, they administered four different types of writing tasks (convincing, describing, instructing, and explaining) to 50 EFL students who were enrolled in the Department of English Language Translation at Islamic Azad University in Iran. The essays collected from the students were assessed by three independent raters using holistic and analytic scoring scales. The results indicated intra- and inter-rater reliability across raters' holistic and analytic scores. According to G-study results examining the persons, raters and tasks variance components, the writing tasks given to the students were at the same difficulty with a relatively small variance component (2.14%). However, another set of statistics showed that students performed better with describing than the remaining task types. The relatively larger variance component due to raters (3.40%) and smaller variance component due to the persons (0.26%) indicated that raters varied in their scores to some extent and students' performances differed among tasks to a small extent.

To put it briefly, essay topics and task types had an impact on the performance of the students and the raters' scoring behaviors (Gebril, 2009; Hamp-Lyons & Mathias, 1994; Jennings et al., 1999; Saeidi & Rashvand Semiyari, 2011; Weigle, 1999). It is believed that a good topic with which the examinees are familiar can produce fair judgements (Gebril, 2009; Weigle, 2002). With this in mind, a single topic, which was familiar to students in terms of their future professions, was chosen in this study. In this way, the researcher eliminated the disadvantages that might stem from a single topic by providing students with an attractive topic on which they could show their performance at a maximum level.

Rater's professional background. Among the studies investigating the factors that influence the reliability of the essay scores, some studies focused on rater's professional background (Elorbany & Huang, 2012; Santos, 1988; Song & Caruso, 1996). These studies demystified the professional background features regarding raters' majors study in particular (Elorbany & Huang, 2012), the science field they come from (Santos, 1988) and the status of English in the job description (Song & Caruso, 1996). The investigation of Santos (1988) aimed to explore 178 professors' reactions to EFL essays. While 96 of them came from departments in the humanities/social sciences, 82 professors represented departments in the physical sciences. The participant professors were asked to rate two EFL essays using a 10-point scales based on two foci—language and content. The results showed that the language component of the essays received higher scores than the content component. The professors in humanities/social sciences tended to score the essays more leniently than did the physical science professors. Additionally, there were two significant variables described for the differences found between professors' ratings in terms of the language but not the content of the essays: age and native language. First, younger professors were found to rate the language more negatively than the older ones. Second, the non-native English speaking (NNS) raters assigned lower ratings to the acceptability of the language, indicating that they found the essays to diverge from the target language norms and features more so than the NES raters did.

In the same vein, Song and Caruso (1996) conducted a study to uncover the differences between the scorings of raters from different professional backgrounds. A number of 30 ESL faculty and 30 English faculty members participated in the research and both groups were asked to rate two ESL and two NES essays holistically and analytically. Raters from the English faculty assigned significantly higher scores to the essays holistically than the ESL faculty raters did. However, no significant differences were found between the rater groups in their analytic scorings. The differences in holistic but not in analytic ratings may have stemmed from the standards that both rater groups built toward English compositions from their professional perspectives, which may have been diminished by the guidance of an analytic rubric.

More specifically, Elorbany and Huang (2012) investigated whether raters' professional background impact ESL writing assessment using a G-theory approach. Twenty teacher candidates with no previous formal teaching experience and who spoke English as their native language participated in the study. Ten TESOL major teacher candidates and ten non-TESOL major teacher candidates scored three ESL essays using a 10-point holistic scale. The results showed that while teacher candidates studying TESOL scored the essays more consistently and reliably, non-TESOL teacher candidates varied considerably in their scores, indicating that the professional background of the raters had an impact on their scorings of ESL essays.

As can be seen in the review of the previous studies, the professional background of raters was found to have an impact on their essay scores assigned to students' essays (Elorbany & Huang, 2012; Santos, 1988; Song & Caruso, 1996). Scoring differences might be related to the raters' expectations and perspectives related to their professional backgrounds and formal training, resulting in issues of rater leniency and severity as well as issues with inter-rater reliability when raters from diverse backgrounds assess the same group of essays. Although the participants included in this study had different levels of education (BA and MA degrees), they came from similar educational backgrounds (English Language Teaching and English Literature) and worked as EFL instructors at Turkish universities, thus minimizing differences that might arise from their professional background.

Rater's L1. As one of the factors affecting ESL/EFL writing assessment, raters' linguistic background has also been under investigation to see whether NES teachers NNS teachers respond to students' writing with similar judgements (Connor-Linton, 1995; Hamp-Lyons & Zhang, 2001; Hinkel, 1994; Hughes & Lascaratou, 1982; James, 1977; Johnson & Lim, 2009; Kobayashi, 1992; Kobayashi & Rinnert, 1996; Santos, 1988; Shi, 2001).

Examining the impact of raters' L1, Kobayashi (1992) investigated NES and NNS Japanese ESL raters' evaluations of ESL essays written by two university-level students. The 145 NES and 124 NNS raters who participated in the study varied in their academic status including undergraduate, graduate and professorial levels. The results showed that NES raters were stricter about the grammar component of the essays than the Japanese native speakers were. Additionally, NES professors and graduate students were more positive in their reactions to the aspects of clarity of meaning and organization than the Japanese speaking groups. However, Japanese undergraduate students assessed the essays more positively than did the NES undergraduate raters. Overall, the aforementioned findings indicated significant differences between the two groups of raters varying in their L1 background.

In the same vein, Shi (2001) examined the differences between 23 NES and 23 Chinese EFL teachers working at tertiary education in China in terms of 10 holistic scores that they assigned to EFL essays and the qualitative reasons for their ratings. The results showed that NES teachers displayed a higher intra-rater consistency than the other group of teachers did. However, NES and Chinese EFL teachers showed no significant difference in their ratings to the students' essays. Although both groups of raters agreed on the most common positive feature—the argument—and the most common negative feature—language, especially intelligibility—of the essays, there was a significant difference between the qualitative judgements of both groups of raters. While NES teachers responded to the content and language aspects of the essays more positively, Chinese EFL teachers attended more negatively to the organization of the ideas and the length of essay. It was suggested that raters belonging to different linguistic backgrounds might base their decisions about ESL/EFL essays on different scoring criteria and qualitative judgements.

More recently, Johnson and Lim (2009) investigated the impact of rater language background on writing performance assessment. A large sample of compositions written by the examinees of Michigan English Language Assessment Battery (MELAB) was rated by 17 MELAB professional raters who all received standardized rater training and completed certification programs. However, while the majority of the raters were native speakers of English, only four raters represented three different L1 backgrounds other than English. In the study, essays written by ESL examinees were scored using a 10-point holistic scale. Employing the IRT FACETS program, the analysis of the data reveled no bias in the ratings stemming from raters' language background. The researcher, however, underlined the lack of generalizability of the findings given the limited number of NNS raters in this study.

As can be seen, the literature is inconclusive as to whether raters' L1 might affect the essay scores significantly (Koyabashi, 1992) or whether the L1 background of the raters might have no observable effect on the essay scores (Johnson & Lim, 2009; Shi, 2001). This factor was not a consideration in this research since all the participants were native speakers of Turkish. In selecting raters with L1 Turkish backgrounds, the researcher intended to eliminate any possible L1 effects in rater variation.

Rating methods. Although using rubrics during writing skill assessment is believed to provide the raters with a sound basis for their scores and interpretations derived from those scores, scoring with a rubric may not make a reasonable difference compared to the criteria-free evaluations unless the raters are trained to use the scales effectively (Rezaei & Lovorn, 2010). However, holistic and analytic scoring rubrics have been used in ESL/EFL writing assessment to identify the students' writing proficiency for different purposes (Bacha, 2001;

Barkaoui, 2007b) given that both rubrics are different in rating methods, assumptions, and implications for essay marking processes and scores given to students' writing (Goulden, 1992, 1994; Weigle, 2002). Holistic scoring rubrics prioritize the strengths of a learner's performance on a writing task; however, analytic rubrics are better for uncovering the learner's weaknesses and are more user-friendly for providing feedback to the weak areas of learners' writing skill (Charney, 1984; Cohen & Manion, 1994; Cumming, 1990; Elbow, 1999, Hamp-Lyons, 1990, 1991, 1995; Reid, 1993; Weigle, 2002; White, 1994). Furthermore, holistic rubrics are considered weak in reliability but strong in validity while multiple-trait scales stand out as a more reliable and practical assessment criterion (Perkins, 1983). Although analytic ratings are thought to be advantageous compared to holistic ratings to assess the quality of L2 writing for the purposes of the assessment such as research, high-stakes testing, or diagnostic assessment, (Charney, 1984; Cohen & Manion, 1994; Cumming, 1990; Elbow, 1999, Hamp-Lyons, 1990; Reid, 1993; Shi, 2001; Weigle, 2002; White, 1994), there have been several studies in the literature investigating the effectiveness of scoring rubrics for attaining high levels of assessment reliability, detecting students' weaknesses in writing ability, and supplying a good amount of feedback for the learners (Bacha, 2001; Goulden, 1992, 1994; Weigle, 2002).

In order to investigate the differences between the scorings assigned holistically and analytically, Song and Caruso (1996) conducted a study with 30 ESL faculty and 30 English faculty members. It was found that raters from the English faculty assigned significantly higher scores to the essays holistically than ESL faculty raters did. However, no significant difference was found between the rater groups in their analytic scorings. This may be related to the features of the scales in that holistic method does not allow an in-depth examination of the essays, resulting in raters' expectations playing a greater role in the holistic scores compared to analytic ratings.

In another study examining differences between rating methods, Bacha (2001) investigated what holistic and analytic scoring methods mean to their users in terms of the evaluation of writing. Using holistic and analytic rubrics respectively, two raters scored a number of thirty essays written by Arabic EFL students at Lebanese American University. The results indicated a high intra- and inter-rater reliability with holistic scores; yet holistic rubrics were insufficient to display students' performance in different components of writing. As for the analytic scoring, the results underlined a positive correlation between the scores given to each component and analytic scores. Additionally, analytic scoring revealed that students performed significantly differently in distinct components of writing from best to least as follows: content, organization, mechanics, vocabulary, and language. Further, significant positive relations were found between holistic and analytic scores. The study suggested the use of a combination of holistic and analytic rubrics to better assess students' writing performance.

Using qualitative and quantitative methods, Barkaoui (2007b) researched the effects of holistic and analytic scoring scales on EFL essay marking processes including essay scores, raters' decision-making strategies, and raters' perceptions of EFL essay scoring in Tunisia. A total number of 32 essays on two argumentative topics written by 16 EFL university students were rated by four EFL teachers holistically and analytically. In addition to quantitative analysis of the ratings within G-theory analysis, TAPs were employed in the ratings of two sets of four essays during holistic and analytic scoring of the essays for the qualitative data analysis. Surprisingly, the results indicated higher inter-rater reliability from the holistic scoring scale than from the analytic scoring scale, a finding contrary to what the researcher had expected; yet, the scores assigned to the component of organization on the multiple-trait scale displayed high reliability as well. Additionally, more decision-making statements were obtained with the holistic scoring scale than the multiple-trait scale. As was expected, the multiple-trait scale resulted in more judgement strategies, and raters used more interpretation strategies with the holistic scoring scale. Regardless of rating method, the findings underscored that the major source of variability in scores and the frequencies of decision-making strategies was the raters.

In a different manner, Knoch (2009) compared two types of analytic rubrics: a previously developed scale with less descriptors and an empirically developed analytic rubric with more descriptors. In doing so, 10 experienced raters were asked to score a total of 100 essays using both of the rubrics. The analysis yielded that raters were more consistent in their scores assigned with the newly developed detailed rubric and qualitative findings showed that raters were more favorable of the more detailed rubric since they were able to better distinguish between the different aspects of writing.

Considering the connection between rater cognition and scoring criteria, Barkaoui (2010c) examined the impact of rater experience and rating methods on the variability of essay scores through data collected from TAPs. Fourteen experienced and 11 inexperienced raters participated in the study and assessed 12 essays both holistically and analytically. The results revealed that rating scale type had a larger effect on raters' decision-making behaviors and the aspects of writing that raters attended to than rater experience did. The results showed that raters' behaviors varied based on the scoring method in that raters attended to the essay itself while using holistic scale, but they referred to the rating scale while evaluating the essays analytically. More judgements strategies than interpretation strategies were employed while using both of the scales overall. However, the holistic scale elicited more interpretation strategies and language focus strategies than the analytic scale, which elicited more judgement strategies and self-monitoring focus strategies.

More recently, Han (2013) examined whether using analytic and holistic scoring rubrics yielded significant results in terms of score variability in one experimental and one natural context in Turkey. In the experimental context, the researcher gave 72 EFL essays to 10 raters who received detailed training prior to scoring. Using a holistic and an analytic scoring rubric, each rater evaluated the essays. In the natural context, nine raters who were oriented but did not receive detailed training scored the same set of essays using the same analytic and holistic rubrics. The data obtained from the experimental context (1,440 ratings) and natural context (1,296 ratings) were analyzed using G-theory approach. The analysis of the experimental context data suggested that following a detailed training, the scores produced using the holistic rubric were as consistent and reliable as the scores produced from the analytic rubric. However, the raters who did not receive the detailed training (the natural context) displayed a great deal of variety in their holistic and analytic ratings. The findings suggested that holistic scoring rubrics could be preferred over analytic scales, even for high-stakes tests, as long as raters are carefully trained with specific consideration of institutional objectives.

To sum up, the type of scoring rubric used is an important factor in attaining consistent scores from raters. While raters tended to differ significantly in their holistic scores in some studies (Song & Caruso, 1996), some studies found that raters varied significantly in their analytic scores and high inter-rater reliability was found in the holistic ratings of scorers (Bacha, 2001; Barkaoui, 2007b; Han, 2013). This study did not aim to compare rating scales and preferred using an analytic scoring rubric given that they serve better for uncovering the learner’s weaknesses and stimulating raters’ thoughts about the aspects of the essay (Charney, 1984; Cohen & Manion, 1994; Cumming, 1990; Elbow, 1999, Hamp-Lyons, 1990, 1991, 1995; Reid, 1993; Weigle, 2002; White, 1994). In this way, the researcher aimed to derive rich interpretations for the qualitative data collected through TAPs.

Rater training. Although a pre-determined scoring rubric is supposed to provide insight to the raters for their evaluations, it may fail to help bring objectivity to the process. Thus, an important way to reduce rater-related reliability risks is to give initial and ongoing rater-trainings to the raters (Barrett, 2001; Lumley & McNamara, 1995; McNamara, 2000; Sweedler-Brown, 1985; Weigle, 1998, 1999; Wolfe, 2005) in which a set of performances at different levels are assessed by the raters independently and discrepancies among judgements are discussed to bring about agreement on the interpretation of the performances (McNamara, 2000). However, other researchers are hesitant to recommend training raters since it may pose a threat to individual approaches to essay-reading by neglecting experience and background for standardization and pushes raters to focus on superficial aspects of compositions (e.g., Barritt, Stock & Clark, 1986; Charney, 1984; Huot, 1990). As such, several studies attempted to assess the benefit of rater trainings with regards to increasing consistency among scorers in the assessment context (Attali, 2015; Shohamy et al., 1992; Sweedler-Brown, 1985; Weigle, 1994).

To this end, Sweedler-Brown (1985) examined the impact of training and experience on the consistency of essay scores at the writing development program of a large university. A number of 26 raters, 6 of whom were highly experienced trainers, participated in this study. While the regular 20 raters received surface training only during the grading session, the 6 trainers were exposed to a very detailed and extensive training prior to the study. Firstly, 897 essays were double-scored by the regular raters holistically and 36 of the essays received scores diverging by more than one scale point; therefore, two raters and one trainer were randomly called to score these essays using the same holistic rubric. Following a certain period of time, these essays were given to the same three scorers to be assessed analytically in order to see any potential effects of training and experience. The results showed that both groups of raters took the same aspects of writing into consideration in their evaluations, which were content and sentence structure. However, scorers with more training and experience showed greater consistency between their holistic and analytic scores, suggesting that training and experience may impact the reliability of the scores positively. Another finding revealed that raters with more experience and training assigned lower scores in both holistic and analytic evaluations, meaning that experienced raters might treat the essays more critically.

Considering the interaction of experience and training, Shohamy et al. (1992) investigated the impact of training and professional background on the reliability of essay scores. In doing so, 20 raters—half of whom were experienced and half of whom were inexperienced—were sub-grouped into trained and untrained raters in the study. Every group of teachers scored a set of 50 essays using three rating scales in their evaluations: holistic scale for general quality, communicative scale for content and argumentation, and accuracy scale for grammar and appropriate vocabulary. The results showed that all groups of teachers, regardless of their professional background and training, displayed high inter-rater reliability. Furthermore, training had a significant impact on the scores while professional background had no effect, suggesting that effective training could bridge the gap between experienced and novice raters.

In the same vein, Weigle (1994) researched the impact of training on experienced and inexperienced raters in the context of ESLPE at an American university. The composition subset of the aforementioned test was the focus of the study, in which 16 raters participated. Using the ELSPE scoring rubric, which included three aspects of content, rhetorical control, and language, raters assessed four essays, two of which were written on a graph prompt while the other two were on choice prompts. All the raters scored the same essays before and after the norming session and data were obtained from interviews conducted at both times along with verbal protocols. The results showed that four of the inexperienced raters assigned different scores to the same essays after the training while the other four inexperienced and eight experienced raters gave similar scores at both times. In accordance with the findings, the interviews and the TAPs of the four least consistent raters were analyzed and it was found that the training the raters received helped them clarify the scoring criteria, revise their expectations from the examinees regarding their level and what they could do, and alleviated inter-rater reliability concerns.

In light of research suggesting the positive influence of training on raters' scoring, Knoch, Read, and Randow (2007) compared the effects of online rater training and face-to-face training in a university-level writing assessment program in New Zealand. A total of 16 raters were placed into two training groups randomly in equal numbers and were asked to score 70 writing samples (five sets of 14) analytically before and after the training phase. The findings indicated that both rater-training methods were effective in terms of severity. There were no differences between the groups when inconsistencies and central tendency effects were analyzed, although online training appeared slightly more effective in reducing differences between raters in terms of leniency and severity. Finally, the qualitative findings suggested that raters would prefer a type of training that combines the two training methods, given that human contact should not be entirely ignored.

To further explore the effectiveness of online training, Attali (2015) investigated the impact of initial web-based and short training programs on the rating performance of inexperienced raters compared to that of expert raters concerning the severity and reliability of essay scores. The holistic scores assigned to 200 essays in the argumentative and issue task forms by 14 inexperienced raters were compared to the data collected in a previous study (Attali, Lewis, & Steier, 2013) in which 16 experienced raters holistically scored the same set of essays. The results indicated no significant differences between both groups of raters with respect to average scores, but significant differences were found between the groups in terms of the variability of scores. Inexperienced raters who received the initial training showed less variability in their average scores although their individual scores were more variable. Furthermore, the results suggested that rater performance was influenced less by actual experience in rating essays while initial training and previous abilities prior to training affected rater performance more.

In light of the aforementioned discussions based on the empirical research, it can be asserted that training raters impacts the essay scoring process positively (Sweedler-Brown, 1985) and can help reduce the variability between experienced and novice raters (Shohamy et al., 1992). Although some researchers were critical of training raters given that it might restrict the personal interpretations of the raters for the assessment task (Barritt et al., 1986; Charney, 1984; Huot, 1990), it should be regarded as an effective way of reducing variability among the raters (Attali, 2015; Barrett, 2001; Han, 2013; Lumley & McNamara, 1995; McNamara, 2000; Sweedler-Brown, 1985; Weigle, 1998, 1999; Wolfe, 2005). The impact of rater training was not investigated in the current research; yet the researcher involved the participants in adapting the scoring scale used to score the essays to minimize the variability that might be caused by a lack of orientation prior to the assessment task. In this sense, although the raters did not receive a detailed training due to practical limitations, the raters were oriented to the analytic scoring scale prior to rating.

Socio-political aspect of writing assessment. Several factors including the test-taker, the prompt or task, the written text itself, the rater(s), and the rating scale that are considered to have an impact on variability of essay scores (Hamp-Lyons, 1990; Kenyon, 1992; McNamara, 1996) have been investigated extensively. However, limited research has been conducted to explore how raters' decisions may vary based on the socio-political aspect of the testing situation. In this regard, Baker (2010, p. 135) considered the test stakes—"relative seriousness of the consequences of a given test score on the test taker and other stakeholders"—as a noteworthy variable that might affect the variability of test scores. The researcher investigated whether raters' assessments of ESL essays varied when the stakes were high or low. Using exploratory mixed-methods research design, three raters scored a sample of 50 to 54 ESL essays written by prospective teachers at a teacher certification program called EETC (The English Exam for Teacher Certification) in Quebec, Canada. As can be understood from the context of the testing, the stakes were very high not only for test-takers but also test developers, administrators, raters, and even proctors. The participating raters were regular raters of EETC and had varying writing assessment expertise. The raters were provided with the same essays to be assessed in the research context four months after the authentic writing assessment. It was assumed that even if the raters felt familiar with some of the essays, they would not remember the scores that they had assigned previously. With the exception of the stakes and assessment conditions (high-stakes in an authentic condition and low-stakes in a research condition), all other variables were held constant in order to attribute any scoring differences to the factor under investigation in this study. The results showed that Rater 1, the most experienced, and Rater 3, the least experienced, assessed the essays in the low stakes condition more leniently than the high-stakes condition. However, the scores that Rater 2 assigned to the essays did not show any significant differences between the two conditions. Furthermore, all the papers that failed in the authentic condition failed in the research condition as well.

The results of the qualitative data collected through post-rating interviews indicated that Rater 2 and Rater 3 felt stricter and less worried about giving lower grades to the essays in the research condition because of the lack of consequences to the test-takers (Baker, 2010). Additionally, they reported a sensation of déjà vu associated with the essays they had previously graded, resulting in some sort of pressure to stay stable in terms of rewarding the similar scores to the same essays. In order to maintain the focus of the study and suppress concerns related to the stability of the ratings, the researcher made use of a fourth rater who was matched with Rater 5 (a pilot rater and not a participant of the research study). Rater 4 was told to score the essays in the research condition. The results showed that Rater 4 assigned exactly the same mean scores to the essays in the low-stakes condition as Rater 5, her matched partner, had done in the high-stakes condition. As a result, Raters 2 and 4/5 were consistent in both assessment conditions while Rater 1 and 3 gave significantly higher scores in low-stakes testing condition.

Based on Baker's (2010) research perspective, the assessment context of this dissertation can be considered low-stakes since raters graded the essays for research purposes. However, the researcher made sure that the participants were aware of the importance of the study to address the assessment problems in EFL writing scorings in Turkey to collect authentic information from their evaluations as much as possible.

Impact of Rater's Rating Experience on ESL/EFL Writing Scores

The expertise and knowledge that each rater brings to the assessment task are essential for the reliability and validity of ratings (Schoonen, Vergeer, & Eiting, 1997). Therefore, researchers in the field have focused on the impact of the previous rating experience of the raters on the various aspects of the scoring process (Barkaoui, 2010c, 2011a; Cumming, 1990; Leckie & Baird 2011; Song & Caruso, 1996; Wolfe et al., 1998).

In 1990, Cumming investigated whether raters implicitly distinguish students' L2 proficiency and writing expertise in their mother tongues while scoring ESL compositions holistically. In doing so, seven novice and six expert raters were given 12 ESL essays, which represented two levels of ESL proficiency (intermediate and advanced), generated by students from two levels of writing expertise (average and professional writers). The results indicated that both groups of raters' evaluations distinguished L2 proficiency and writing expertise as non-interacting separate factors. Additionally, novice and expert raters significantly differed from each other in their ratings of 'content' and 'rhetorical organization' but the ratings assigned to 'language use' did not differ significantly. Statistical analysis of the ratings also showed that expert raters scored more consistently than novice raters with regards to the scores given to three aforementioned aspects of the compositions. Moreover, the analysis of qualitative data collected through TAPs reveled 28 common decision-making behaviors displayed by novice and expert raters. Both groups of raters varied significantly in terms of employing different decision-making strategies.

In Song and Caruso's study (1996), 30 ESL faculty and 30 English faculty members with varying previous rating experience scored four essays using both holistic and analytic rubrics. Raters form English faculty were almost twice more experienced than ESL faculty raters and they assigned significantly higher scores to the essays holistically than ESL faculty raters did. However, they did not differ in their analytic ratings significantly. According to the researchers, the lenient behaviors that more experienced raters displayed in their scorings might be related to more realistic expectations they set when their experience in teaching and rating increased.

In order to find out the differences in rating behaviors across different scorers with varying assessment proficiency, Wolfe et al. (1998) examined 36 raters regarding their behavioral differences in relation with scoring proficiency. The participants were asked to score 24 essays holistically while thinking aloud. Then, two independent individuals who had experience in writing assessment research and verbal report analysis coded the protocols and categorized the statements into 4 groups including the essay feature referenced, the degree of specificity of the statement, the degree of rubric adoption demonstrated, and the cognitive task performed. The results indicated that the most experienced raters did not differ in the scoring foci to which they attended; however, the least experienced scorers put slightly more emphasis on storytelling than the other foci, mechanics, organization, and style. In terms of degree of specificity, more experienced raters were more likely to make general comments in their evaluations while scorers with less proficiency tended to focus on specific features of the essays. More experienced raters articulated rubric-generated utterances in their discussions, whereas respectively inexperienced raters were keener on relying on self-generated statements based on the particular features of the essays, suggesting that more experienced raters may tend to base their assessments on task-specific rubrics rather than using intuitive assessment criteria independently as did less experienced scorers. Finally, the researchers found that more experienced raters used a more holistic approach while assessing the essays in that they made their judgements by reading and evaluating the whole essay and assigning an overall score; on the contrary, less experienced raters tended to adopt a bottom-up approach and evaluated the essays with a step-by-step read-assess strategy.

In 2001, Rinnert and Kobayashi probed the perceptions of four groups of raters with different levels of experience in scoring writing. The study included a number of 465 raters, 106 of whom were NES teachers and served as a control group while the remainder came from Japanese backgrounds (127 inexperienced EFL students, 128 experienced EFL students, 104 Japanese EFL teachers). They evaluated the essays produced by Japanese EFL students by assigning scores to six criteria including clarity of meaning, quality of content, quality of introduction, quality of conclusion, logical connection of ideas, and language use as well overall quality on 10-point scales. The results showed that the four groups of raters differed significantly with regards to their scores for all the criteria except quality of conclusion. However, the least experienced Japanese EFL raters differed most from the NES rater group in that the former assigned the highest scores to the essays, whilst the latter group of teachers gave the lowest scores. Furthermore, the criteria that the raters attended to in their evaluations varied among the groups in that inexperienced student raters put the greatest emphasis on the content whereas experienced student raters and Japanese EFL teacher prioritized logical connection and clarity in addition to content and, as for the NES raters, clarity showed the highest correlation with their judgements for the overall quality. Finally, the qualitative analysis of the raters’ comments on the compositions revealed that experienced Japanese EFL raters and NES raters shared similar perceptions of the texts, following a pattern of preferences that tended to change gradually from L1 writing features to those of L2 writing as the experience increased.

Furthermore, Wolfe (2005) reanalyzed the findings of Wolfe et al.'s (1998) study from additional two more perspectives—*jumps* and *hits*. Firstly, he investigated the degree to which the raters with different proficiency levels shift their attention across categories in the rating process. Although less experienced raters were expected to employ a read-monitor-read-monitor reading style during their evaluations that may end up with a more frequent content focus category jump behaviors, the findings did not suggest any significant differences between proficiency groups in terms of jumping between categories. Secondly, the study examined whether there were any differences between raters with varying scoring experience in terms of the number of categories they mentioned in their assessments. In this respect, proficient raters were hypothesized to cover a greater number of categories while making decisions about the essay because of their previous expertise in using scoring scales. Yet, no meaningful difference was found between proficiency groups regarding the number of categories hit.

Considering the connection between raters' experience and their use of different scoring scales, Barkaoui (2010a) examined the holistic and analytic scores that 31 novice and 29 experienced raters assigned to a total of 24 ESL essays to see the effects of different marking methods and rating experience on essay scores in terms of inter-rater agreement, rater severity and self-consistency. The results showed that both marking methods measured the same constructs in the essays. However, a higher level of inter-rater agreement was observed in holistic scores while there was a higher self-consistency with analytic marking. Both rater groups were more lenient when assessing essays analytically; novice raters were more lenient while using both of the scoring methods, though. While the analytic scoring method resulted in within group variability and severity, the differences in average severity were larger with the holistic scoring method. Moreover, novice group of raters displayed more inter- and intra-rater variation in terms of severity. The findings showed that while marking method influenced rater self-consistency, rating experience influenced inter-rater reliability. Finally, the researcher suggested that different scoring methods might be required for different assessment purposes, contexts, and raters, and he further claimed that analytic scoring methods might be more appropriate for less experienced raters given that the rubric can canalize their motivation and attention to the marking criteria and can enhance their self-consistency.

In a large-scale assessment context, Leckie and Baird (2011) examined rater effects with regards to rater's severity drift, central tendency, and their previous rating experience among three rater groups including team leaders, experienced raters' and new raters by analyzing their scores assigned to England's national curriculum English writing test for 14-year-old students. In this test, the students were asked to answer two essay questions; however, this study focused on the ratings of one essay question that accounted for 60% of the total points. A total number of 34,920 ratings given by 135 team leaders, 372 experienced raters, and 182 new raters were analyzed to see whether rater severity and rater central tendency varied based on prior rating experience. The experienced raters had at least one year more scoring background in the aforementioned national-level test context. The results showed that raters' levels of severity did not differ significantly over time and the raters became more homogenous when they scored more essays. Nevertheless, significant intra-rater variability in rating severity was found over time. In general, raters over-scored low-quality essays and underscored high-quality essays, resulting in raters' central tendency. As for the raters' previous rating experience, it was found that experienced and new raters did not score significantly different in terms of their rating severity. Additionally, their scores did not significantly deviate from the consensus scores assigned by the expert committee. However, being the most experienced raters, team leaders, who were monitoring and checking the experienced and new raters' scores at certain intervals, significantly over-scored by half a point out of the maximum score (30 points) compared to the expert committee. These findings suggested that as raters become more experienced, they might seek different criterion aspects, which are not included in the scoring rubric.

Considering experience from a developmental perspective, Lim (2011) conducted a longitudinal study to examine to what extent experienced and new raters' rating quality change over time by using the scores assigned to the writing section of the Michigan English Language Assessment Battery (MELAB). The research was conducted in three time periods and included a mixture of 11 experienced and new raters. In doing so, two new raters were added to the design in each time period, resulting in the involvement of 6 novice raters in total. The results indicated that the experienced raters' scoring quality stayed the same from the beginning to the end of the study, whilst novice raters improved their rating quality upon practice, suggesting a positive correlation between rating volume and rating quality. Overall, the findings of this research indicated the importance of rating practice on the moderation of the novice raters in the long run in terms of acquiring expertise in writing assessment.

To summarize, the previous rating experience that raters have has an impact on the essay scores and rating behaviors exhibited during the assessment task. It was found that experienced raters performed more consistently while assigning scores than novice raters (Cumming, 1990; Lim, 2011). In addition, the performance of raters with varying experience changed depending on the rubric used for evaluating the compositions in that more experienced raters assigned higher scores holistically compared to raters with less experience, although the two experience groups did not differ significantly with respect to their analytic ratings (Song & Caruso, 1996). However, in other studies less experienced raters were found to give higher scores to essays (Barkaoui, 2010a; Rinnert & Koyabashi, 2001). Building on previous research, this study compared the scores that raters with varying experience assigned to the EFL essays. In doing so, instead of making a clear-cut distinction between inexperienced and experienced raters, in this study essay scores were compared among three experience groups in accordance with raters' self-reported previous rating experience.

Impact of Essay Quality on ESL/EFL Writing Scores

L2 proficiency and expertise in writing are different but not unrelated (Cumming, 1989; Krapels, 1990; Kroll, 1990) in that students can benefit from their command in writing in their L1 while generating a text in an L2; however, this production process can be hindered because of the priority to focus on language (form) rather than the content (message) (Weigle, 2002). In order to support this view, Weigle (2002) mentions Hayes' (1996) model to explain the cognitive process through which L2 writers go through and the difficulties they experience in text interpretation and text generation. In addition to the English language proficiency of the students, their L1, home culture, and style of written communication can be listed as factors that can affect ESL students writing in terms of paper quality (Hinkel, 2003; Yang, 2001). These factors can also influence the behaviors that raters exhibit during ESL writing assessments (Bachman, 2000).

Another remarkable point is the contrast effect in rating while assessing papers of different qualities simultaneously. While a medium quality essay tends to receive a low score when it is assessed after reading several high-quality essays, it tends to receive a higher score when it is preceded by a number of lower quality essays (Daly & Dickson-Markman, 1982; Freedman, 1981; Hughes & Keeling, 1984). In addition to this, rater expectation is another aspect that is worthy of discussion in terms of its impact on essay scores because raters tend to assign higher scores to the same essays when they are told that the essays are written by better students (Diederich, 1974). In light of these discussions, the impact of text quality has been in the focus of writing assessment reliability studies (Brown, 1991; Engber, 1995; Ferris, 1994; Han, 2017; Huang, 2008; Huang et al., 2014).

Considering the impact of text quality, Brown (1991) investigated differences between the scores assigned to compositions written by international students and NES students. In doing so, 112 compositions were collected to be scored holistically by eight raters who were from ESL and English faculties at the University of Hawaii. Before scoring the essays, raters were trained to use the rubric and students' essays were labeled with numbers in order to avoid any rater biases that could arise from students' backgrounds. The results suggested no significant differences between the ESL and NES students' compositions and the ratings given by ESL and English faculty members. Furthermore, content and syntax were considered the best and the worst features of compositions respectively. However, the raters showed differences in other features of analysis such as cohesion, content, mechanics, organization, syntax, and vocabulary, suggesting that the raters might have arrived at their scorings from different points of view. Thus, the results suggest the importance of the decision-making process in assessment research.

Investigating the lexical and syntactic features of compositions written by ESL students at different proficiency levels, Ferris (1994) benefited from a corpus of 160 ESL essays to identify quantitative, lexical, and syntactic features, resulting in 28 text variables used for the statistical analysis. Three independent raters graded the essays holistically and the scores were used to place students coming from different L1 backgrounds into proficiency levels. Following the aforementioned method, students were grouped into lower level groups and advanced levels. While the lower level group had a mean score of 14.8, the mean essay score for the advanced group came to be 22.9 out of 30. Further analysis of the texts showed that the 28 text variables discriminated the students into level groups with 82% accuracy. Additionally, 18 of the variables differed significantly between groups. These findings showed that students with higher proficiency employed textual features in target language writing compared to lower level students. Moreover, they were able to use more diverse syntactic and lexical tools while writing ESL compositions, suggesting implications for writing instructors to help students develop their writing strategies by enriching their lexical choices, rhetorical patterns, and syntactic constructions to receive higher scores.

Similarly, Engber (1995) investigated how the lexical component is related to the quality of compositions written by ESL students from various L1 backgrounds. The data for this study included 66 essays and student writers were enrolled in four distinct proficiency levels in an intensive English program at a university in the United States. Ten raters scored the essays using a 6-point holistic scoring scale and statistical analysis revealed a high inter-rater reliability (*r* = .93). Average scores per essay ranged from 1.6 to 5.6. Following that, the errors related to the lexical component was analyzed and categorized based on four lexical richness measures including lexical variation with error, lexical variation without error, percentage of lexical error, and lexical density. The results showed that the scores assigned to the essays decreased when the lexical errors increased. However, lexical error or lexical variation alone were not enough to explain the quality scores that raters assigned to the essays. That is to say, higher scores were assigned to essays in which lexical variety was used correctly.

Considering text quality on the basis of language proficiency, Huang (2008) examined the reliability of scores assigned to the essays produced by ESL students and NES students in large-scale provincial English examination in the years of 2002, 2003, and 2004 in Canada. The students were asked to respond to three types of tasks such as writing a unified and coherent paragraph about a poem, a multi-paragraph write-up about a literary prose, and producing a multi-paragraph original essay. The results showed that ESL students' performance was significantly lower than that of the NES students across three tasks and three years. The difference found between the groups were attributed to one of or a combination of the following factors including the possibility that ESL students might have difficulty understanding the tasks, rating bias against ESL students, and the fact that ESL students varied systematically in their writing skills. However, the greatest factor contributing to the variability of scores between the two groups was found to be English writing skills of the students. These results indicated a fairness problem between the scorings of ESL and NES students.

Using G-theory, Huang et al. (2014) investigated the impact of essay quality on rating variability and reliability of ESOL writing at a Turkish university. Five ESOL raters scored nine argumentative essays in three distinct qualities including low, medium, and high with holistic and analytic scoring scales. The raters did not receive any formal rater training prior to this study; yet, they were calibrated to both rubrics by rating five essays of different qualities before the main data collection. According to the results, holistic scoring method resulted in greater standard deviations for low- and high-quality papers while a smaller standard deviation was observed for the papers in medium-quality. Analytic scores yielded higher mean scores compared to the ones obtained holistically. Additionally, the participant raters scored high-quality essays more consistently and displayed more variety in their scores assigned to low-quality essays. Furthermore, scoring method contributed to the scores of high-quality papers, whereas it did not affect the scoring of low-quality essays. Overall, the findings underlined that the quality of the essays affected the raters' holistic and analytic scores considerably.

In a recent study, Han (2017) examined the holistic scores that raters assigned to EFL essays of different qualities—low, medium, and high. The essays were collected from three universities and 30 essays of distinct qualities were obtained and used in the research. Five volunteer raters scored these essays using a holistic scoring scale. In addition, the raters were asked to implement TAPs while assessing six essays, two from each of the three categories, to examine their decision-making behaviors toward the essays of different qualities. The results revealed that the raters assigned similar scores to high-quality papers compared to the low-quality papers. Furthermore, G-theory analysis showed that the largest variance component (37.2% of the total variance) was found to be essay quality, followed by rater (24.8%), indicating that papers fluctuated greatly in their quality and raters displayed remarkable differences in their assigned scores.

All in all, raters exhibit different scoring behaviors and the score variability exhibits different patterns when considering essays of different qualities determined by the author students' L1 (Brown, 1991; Huang, 2008) or the students general command in writing (Han, 2017; Huang et al., 2014). Given these points, this study inquired into the score variability between essays of two distinct qualities, thinking that different rating behaviors might occur while assessing high-quality and low-quality papers with specific focus on the interaction between rater experience and assigned scores to essays of different qualities.

Rater Cognition and Decision Making While Rating

Given the complexity of writing skill, scoring scales alone cannot capture the multifarious nature of aspects such as grammar, content, lexical usage, and coherence into simple scale points. As such, they may not be sufficient to understand the essential characteristics of students' writing performance and may hinder the rich and multi-faceted interpretations of human raters (Cumming, Kantor & Powers, 2002; Henning, 1991; Raimes, 1990). Therefore, understanding a rater's cognition and how it relates to that rater's decision-making process is important (Vaughan, 1991); in particular, individual characteristics such as experience and proficiency may be fundamental to writing assessment research (Baker, 2012). DeRemer (1998) asserts that raters should not simply be treated as a bridge between the text and the scoring criteria but rather it should be noted that they engage in a constructive operation akin to a problem-solving activity while evaluating an essay. Furthermore, she defines writing assessment as an ill-structured task in that there is no standard solution for assessment problems despite standardized training procedures. In this regard, even when experienced raters are trained to use specific scoring criteria, they display great variability in their behaviors that are characterized with various reading styles comprised of rater-specific ways to focus on and process the information relevant to the essays (Eckes, 2008, 2012). The different opinions that the scorers have may indicate that they think dissimilarly about the distinct features of an essay. That is to say, they rely their assessments on their individual beliefs and opinions about the essay, resulting in a potential source of error (Wolfe et al., 1998).

In order to comprehend what goes on in raters' minds, researchers developed models to represent the scorer thinking process systematically (Frederiksen, 1992; Freedman & Calfee, 1983; Wolfe & Feltovich, 1994). In their information-processing model of essay scoring, Freedman and Calfee (1983) focus on three processes that are central to assessing students' compositions: a) reading text to construct a text image, b) analyzing the text image, and c) uttering the evaluation. This model suggests that information is taken from the students' writing from which the scorer creates a text image. Considering the different beliefs, values, world knowledge, and understanding of writing that scorers have and the environmental factors impacting scorers' text-reading processes, the text image is not thought to be the exact reflection of the original writing and can be constructed differently by different raters. Based on the text image, scorers evaluate the essay from various perspectives within their internalized or pre-determined scoring criteria to arrive at a decision about the composition. As detailed by Wolfe (2005, p. 40), in this model, the scorers actually go through a series of mental processes as follows:

- *reading* the text to formulate a text image
- *commenting* on the content without a non-evaluative manner
- *monitoring* the particular aspects of the text to see the extent to which the essay exemplifies the criteria in the rubric while evaluating the created text image

- *reviewing* the most notable features of the essay after reading the text
- *making a decision* about the score
- *rationalizing* the assigned score to justify the decision
- *diagnosing* how the text can possibly be improved
- *comparing* the essay to the other writings in the same set

(Wolfe, 2005, p. 40)

Presenting a different model, Frederiksen (1992) suggests that scorers focus on different scoring foci that are their internalized representations of the scoring criteria to draw conclusions. With regards to processing actions, in this model scorers adopt a bottom-up approach in which they separate the performance into pieces to process several evaluations before making their final judgements. Frederiksen's conceptualization is in contrast to Freedman and Calfee's (1983) model which describes a linear approach in which the raters arrive a scoring decision based on a holistic text image. The differences that occur with respect to how the aforementioned frameworks manifest may be related to the rating proficiency of the scorers (Wolfe, 2005).

Having built upon the previous models, Wolfe and Feltovich (1994) put forward a model to map the complicated decision-making process of scoring in which two primary interpretative frameworks are applied—a model of performance and a model of scoring. The former deals with the characteristics that indicate writing proficiency. The researchers mainly identify four components that the raters may establish in their models of performance including *development* (writing down a story with its details and supporting ideas), *organization* (sequencing ideas and events in a logical order), *voice* (providing insight or display personal style), and *mechanics* (effective use of spelling, punctuation, capitalization, etc.), which correspond to the elements of a scoring rubric. In addition, they describe three more categories containing *appearance* (the textual appearance of the essay), *subject* (compliance with the prompt), and *non-specific* (general comments about the essay). These seven categories are named as content focus and considered as variables that constitute raters' models of performance (p. 15).

The latter interpretive framework, *model of scoring*, is the cognitive representation of a set of processes through which the rater interprets the essay and assigns a score. The *model of scoring* model involves a series of elements that come out with the employment of several *processing actions* as in the following (Wolfe & Feltovich, 1994, p. 18):

Model of Scoring		*Processing Actions*
• interpretation	→	read
• evaluation	→	decision, monitor, review
• justification	→	compare, diagnose, rationale
• document	→	record, change, organize
• interaction	→	comment

(Wolfe & Feltovich, 1994, p. 18)

What makes Wolfe and Feltovich's (1994) model of rater cognition different from the previous model proposed by Freedman and Calfee (1983) is the inclusion of documentation, which is a system that scorers create to record their comments especially in large-scale assessment contexts. Furthermore, this model integrates interaction as the fifth component into the model to refer to the personal involvement of the raters in the reading process.

In addition to the aforementioned studies that focused on rater cognition and decision-making strategies, Baker (2012) synthesized decision-making styles (DMS) benefiting from the relevant literature (Gambetti, Fabbri, Bensi, & Tonetti, 2008; Scott & Bruce, 1995; Spicer & Sadler-Smith, 2005):

Rational DMS: preference for the systematic collection, evaluation, or weighing of information.

> *Intuitive DMS:* preference for relying on feelings, hunches, and impressions that cannot be put into words when making decisions.
>
> *Dependent DMS:* preference for drawing on the opinions or support of others; on receiving second opinions or advice.
>
> *Avoidant DMS:* preference for delaying decision-making, hesitating, or making attempts to avoid decision-making altogether.
>
> *Spontaneous DMS*: preference for coming to a decision immediately or as early as possible. (Baker, 2012, p. 227)

In light of the discussions on scorers' cognitive complexity and the frameworks and models developed to understand how they process essay-relevant information before arriving at a decision, previous research has also focused on the cognitive structures of the raters to investigate their decision-making behaviors while rating (Baker, 2012; Barkaoui, 2010c, 2011b; Cumming et al., 2001, 2002; DeRemer, 1998; Eckes, 2008; Pula & Huot, 1993; Sakyi, 2000, 2003; Vaughan, 1991; Wolfe & Feltovich, 1994; Wolfe et al., 1998). Vaughan (1991) examined the thinking process of nine experienced raters using TAPs. The raters were asked to score six essays using a holistic scale and tape-record the complete scoring process. The transcribed verbal protocols revealed that the most frequently made comments by the raters underlined the weak or unclear content, followed by poor handwriting. In the study five reading styles were identified including *single-focus approach, first impression dominates approach*, *two category strategy*, *the laughing rater*, and *the grammar oriented rater*. Finally, the researcher underscored that:

> [T]he raters are not tabula rasa, and do not, like computers, internalize a predetermined grid that they apply uniformly to every essay. Despite their training, different raters focus on different essay elements and perhaps have individual approaches to reading essays. (Vaughan, 1991, p. 120)

Presenting a rater cognition model, Wolfe and Feltovich (1994) investigated how raters diverged in their models of performance—content focus for judging performance and models of scoring—processing actions to score an essay by designing two studies. In Study 1, six novice and five experienced raters were trained to use a holistic scoring rubric prior to a three-day task of scoring a large number of students' essays from a national essay examination. Then, they were asked to define the characteristics of the papers based on the rubric. In Study 2, six experienced raters evaluated the essays by employing TAPs and, to compare the findings, the scorers were divided into two groups as better and poorer based on their scoring performance. According to the results, four main conclusions were drawn: 1) the thinking process of the raters while scoring the essays were formed by the criteria on which they focused. The models of performance most commonly called upon by the raters in both studies were the development of ideas, organization of content, and the writer's voice. Moreover, the better scorers were more consistent in their use of content categories while discussing the specific characteristics of the papers. 2) The more raters practiced scoring, the more cohesive and complex their models of performance became, suggesting that novice raters can focus on similar content with the expert raters in case where they receive enough practice. 3) The scorers tended to use a model of scoring that contained three moves including reading the essay to interpret the content, monitoring or reviewing the content to decide on the quality, and justifying their decisions by rationalizing the assigned score. 4) Better scorers differed from the poorer ones in terms of models of scoring used in the initial stages. While poorer scorers were more likely to read the whole essay and rarely intervene to comment on the content preceding their evaluations, experienced raters dealt with the text several times by monitoring the content. Additionally, better scorers made non-evaluative comments more, indicating higher degree of interaction with the text.

Four years later, DeRemer (1998) carried out a similar study to investigate three highly experienced raters with a focus on how they defined the assessment task by analyzing differences in strategy usage among raters. Two of the raters were teachers of the students whose essays were assessed within the research context while the other participant was an external scorer. While thinking-aloud, each rater was tasked to score 24 essays chosen from the writing portfolios of eight students. The coding of the verbal reports revealed that raters displayed several operations including rater goals, evaluations, and relations. Additionally, the results showed that three types of task elaboration were derived from the coded verbal-reports, which are search process, simple recognition elaboration, and complex recognition elaboration. The first type—search elaboration—emerged when the rater went through the rubric to find a match between their reaction to the text and the language used in the rubric prior to score assignment and eliminating the alternative score(s). Simple recognition elaboration, however, was present during the time when a score was assigned based on a general impression without any consideration of the criterion being evaluated. On the other hand, complex recognition elaboration came up when the raters scored the essay followed by an analysis of the criteria and the assigned score was justified by relation and evaluation operations. The findings suggested that the different task elaborations evident in this research had different foci, implying that although the raters evaluated the same essays, the scores they assigned did not have the same meaning.

In 2002, Cumming et al. reported a three-coordinated study that aimed to develop a framework to describe the decision-making behaviors of the experienced raters while rating ESL/EFL essays. In Study 1, ten ESL/EFL raters with extensive experience in teaching and assessing writing were employed as both participants and researchers in the study for collecting and analyzing data to develop a preliminary descriptive framework from the verbal protocols that they produced while rating 60 TOEFL essays written on four different essay topics. Most of the raters considered that their scores were influenced by their prior experience; whereas two raters did not report such an influence at all. The authors thought that previous scoring criteria may have affected the raters and it may be quite difficult for experienced raters to change their rating behaviors that they had formed from previous rating experiences. Additionally, the qualitative data results revealed that more experienced raters tended to produce longer and more detailed TAPs and 35 distinct decision-making behaviors were defined. While some raters focused on reading and interpretation strategies more, others preferred more judgement strategies. However, there was not a specific difference among decision-making behaviors across different tasks.

Study 2 of Cumming et al.'s (2002) research included seven highly experienced NES raters each of whom graded 40 TOEFL essays from the same essay pool. The research group comprised in Study 1 analyzed the data obtained from this study and the findings were compared to those acquired in Study 1. According to the raters, their previous experience on writing assessment influenced their ratings in this research. The comparison of the data pertaining to Study 1 and 2 indicated that the seven NES composition raters exhibited basically the same range of decision-making behaviors as the 10 ESL/EFL composition raters did. However, several differences were observed between the two groups of raters as follows: a) NES essay raters assessed the essays after they read them and adopted a cumulative approach to the rating task to bridge their impressions and judgements while the ESL/EFL essay raters followed a progressive pattern through which they made step-by-step decisions while they read the essays, b) NES essay raters tended to evaluate the essays more quickly, reflectively, and creatively, c) While ESL/EFL and NES composition raters paid approximately the same amount of attention to interpreting (40%) and judging (60%) in general, they differed in terms of attention they devoted to decision-making behaviors including self-monitoring (ESL/EFL *M* = 44%, NES *M* = 38%), behaviors pertaining to rhetoric and ideas (ESL/EFL *M* = 19.6%, NES

M = 33.6%), and behaviors related to language components (ESL/EFL *M* = 36.4%, NES *M* = 28.3%).

When further elaborating on the findings, the authors inferred that NES essay raters distributed their attention to points on rhetoric and ideas and to points on language evenly while rating the essays, whereas ESL/EFL essay raters devoted more attention to issues about language than to matters of rhetoric and ideas. Additionally, both NES and ESL/EFL essay raters were found to be pay more attention to rhetoric and ideas in high-quality essays while they devoted more attention to language when they rated low-quality papers.

Study 3 was designed to answer whether different writing tasks would evoke decision-making behaviors in similar qualities, frequencies, and distributions as TOEFL essays did when the same group of raters from Study 1 rated the essays. Six experienced ESL/EFL raters from Study 1 and one additional rater with similar background were asked to score 36 compositions—five separate new tasks and one standard TOEFL essay—produced by six students. The raters were again asked whether their prior experience impacted their ratings and they all answered in an affirmative way. The results indicated that the raters displayed similar decision-making behaviors while rating essays on different tasks; nevertheless, the qualitative data showed that the raters went through deeper and expanded considerations of the prompts because of their complexity compared to TOEFL essays. Consequently, the findings of the three coordinated studies mentioned above revealed that NES composition raters and ESL/EFL composition raters may react to the essays similarly in terms of their decision-making and the attention they devoted to the matters of rhetoric and ideas and language. Additionally, the quality of the essays was found to have an impact on raters' decision-making preferences. Finally, the authors suggested that although the experienced raters exhibited similar decision-making behaviors across different tasks, they may need scoring scales designed for individual tasks given that assessing complex tasks like in Study 3 requires more explicit guidelines.

In the same year, Lumley (2002) investigated the decision-making processes that four experienced raters went through to evaluate two sets of 24 essays written within the context of the Special Test of English Proficiency applied to make immigration decisions in Australia. While the first set of essays were scored analytically without thinking aloud, the raters practiced TAPs during the assessment of the second set. The data obtained from verbal-protocols showed that three general types of behaviors including management, reading, and rating appeared. Further analysis of the data demonstrated that the stages, focus, and behaviors that the raters engaged in overlapped with the findings of Freedman and Calfee (1983). In the stage of reading, the raters attended to global and local features of the text to build a general impression. In the second stage, raters scored the text considering the components of the rubric including task fulfillment and appropriacy, conventions of presentation, cohesion and organization, and grammatical control. In the final stage, the raters reviewed the scores assigned to the texts to confirm their decisions. Additionally, the findings revealed that the scale used in the study did not provide a comprehensive framework to the raters, resulting in developing a variety of strategies to deal with the challenges they faced in the rating process.

Adopting a quantitative approach to cluster rater types in writing assessment, Eckes (2008) examined 64 raters who had expertise in writing assessment within the context of the Test of German as a Foreign Language (*Test Deutsch als FreMdnsprache,* TestDaF). This test is administered to international students who apply for reading universities in Germany and the raters use a rubric that includes 36 different descriptors within a set of criteria across distinctive performance levels. The raters were given a questionnaire asking them to rate the degree of importance that they would attach to each item in an assessment situation. The list of criteria provided in the questionnaire were quite similar to the aspects covered by the TestDaF scoring scale, which were fluency, train of thought, structure, completeness, description, argumentation, syntax, vocabulary, and correctness. The results indicated that raters fell into six types as suggested by the results of the four-point importance scale. Four of the rater types, however, came out as extremely important as follows: the syntax type, the correctness type, the structure type, and the fluency type while the remaining two types were determined by scoring criteria to which raters attached respectively less importance: the non-fluency type and the non-argumentation type. As a result, the findings suggested that rater trainings can be revisited and revised for the raters with different scoring profiles as to redirect their attention to the criteria they may ignore in the scale.

Using the coding scheme developed by Cumming et al. (2002), Barkaoui (2007b) examined the ratings of four raters assigned holistically and analytically to 32 essays on two argumentative topics written by 16 EFL university students. In addition to quantitative analysis of the ratings within G-theory analysis, TAPs were employed in the ratings of two sets of four essays during holistic and analytic scoring. The results showed that more decision-making statements were obtained with the holistic scoring scale than multiple-trait scale. On the other hand, as was expected, the multiple-trait scale resulted in more judgement strategies, while raters stated more interpretation strategies with the holistic scoring scale. However, the rubrics did not affect the rating process markedly regarding the aspects of essays that raters attended to except for one strategy, “read or reread text,” which was employed significantly more while using the holistic scale.

Employing the same coding system, Barkaoui (2010c) investigated the impact of rater experience and rating methods on the variability of essay scores along with examining their interactions through TAPs. Fourteen experienced raters with a minimum experience of five years assessing writing and 11 inexperienced raters participated in the study and assessed 12 essays both holistically and analytically. The results revealed that rating scale type had a larger effect on raters’ decision-making behaviors and the aspects of writing that raters attended to than rater experience did. Furthermore, raters’ behaviors varied based on the scoring method in that raters attended to the essay itself while using the holistic scale, although they referred to the rating scale while evaluating the essays analytically.

In a recent study, Baker (2012) aimed to investigate the impact of individual differences in cognitive style on rater behavior, which had been the focus of previous studies. In doing so, the researcher collected data from six experienced raters through self-report measures, write-aloud protocols, instances of deferred scores as well as scores assigned to 54 papers written within the context of English Exam for Teacher Certification (EETC) in Quebec. The results showed that the most commonly articulated comments were rational (171) and intuitive (129) while other types of comments were made less often, such as spontaneous (72), dependent (29, and avoidant (13). As for the incidents of deferred (doubled) scores, two raters did not use double scores while 20% of one scorer's scores were doubled. The remaining three scorers preferred doubled scores less often with percentages of 18%, 9%, and 4%. Finally, when the combination of all data sources was considered, the most dominant decision-making comments were found to be rational and intuitive (three scorers each). Two scorers used each of the dependent and avoidant styles, whereas only one scorer employed spontaneous decision-making style.

More recently, Han (2017) investigated raters' decision-making behaviors while assessing EFL essays of different qualities. In doing so, he employed the coding scheme developed by Cumming, Kantor, and Powers (2001). Five raters assessed six essays (two from each quality: low, medium, and high) while thinking-aloud; however, the raters were not informed about the paper quality beforehand. The results showed that raters exhibited behaviors related to self-monitoring focus and rhetorical and ideational focus more while assessing low-quality essays compared to mid-range and high-quality compositions. However, raters displayed more language-related behaviors while evaluating mid-range and high-quality papers than they did for low-quality essays.

The above literature shows that raters' decision-making behaviors are connected to several factors including raters' professional background, rating experience, and the quality of the papers. Different models investigating raters' thinking process have developed upon each other, and empirical research has revealed varying findings related to the exhibited decision-making processes of the scorers. Although several studies have been conducted to explore rater cognition over three decades, the thinking processes used by raters might be related to yet unexplored factors such as L1, culture, and the personal characteristics of the raters, which encouraged this research to expand on the thinking processes of the raters during their assessments.

Summary and Research Gaps in EFL Writing Assessment

This chapter touched upon important considerations of reliability and fairness issues in writing assessment followed by a brief summary of the writing assessment situation in Turkish higher education contexts. In later sections, the factors affecting the variability of EFL/ESL writing scores were scrutinized with several empirical research reviews. The chapter continued with a detailed examination of the factors—raters' professional experience as EFL assessors and the essay quality—contributing to the reliability of essay scores; it then elaborated into rater cognition and decision-making behaviors while assessing writing.

Although a multitude body of research has been conducted to investigate the impact of rating experience on the variability of essay scores, the conflicting findings suggest that more experience does not necessarily ensure reliable scores in writing assessments. In addition, very little research has been conducted to investigate issues related to EFL writing assessment in the Turkish context. This study will aim to fill a research gap by investigating rater reliability issues in EFL writing assessment in the higher education context. The findings will provide implications for assessment practices and protocols especially at the institutional level. In addition, the impact of essay quality on the variability of essay scores has been under-researched; therefore, this study will shed light on the interaction between raters and essay quality regarding the essay scores and decision-making processes.

The research in EFL writing assessment has been conducted using different methodological approaches including quantitative theoretical frameworks including classical test theory (CTT), item response theory (IRT), G-theory and qualitative methods such as interviews, write-aloud and think-aloud protocols. However, verbal protocols have not been widely used given their challenging nature in the processes of data collection, preparation (transcribing process), and analysis. With this in mind, this research will contribute to the field by examining rater cognition using verbal protocols. Further, to the researcher’s knowledge, raters’ decision-making behaviors while scoring EFL essays have not been investigated extensively in the Turkish context, a research gap this study aims to fill.

Chapter III

Methodology

The purpose of this thesis study is to investigate the impact of scoring experience of the raters and the quality of the essays on the variability of EFL essay scores and rating behaviors exhibited in Turkish tertiary-level education. Employing a mixed-methods research design, the data for the study were collected both qualitatively and quantitatively. Dörnyei (2007) discusses the strengths and weaknesses of combining quantitative and qualitative methods and states that "[m]ixed-methods research has a unique potential to produce evidence for the validity of research outcomes through the convergence and corroboration of the findings" (p. 45). In the same manner, Mackey and Gass (2005, p. 181) underscore the weakness of using one method in terms of providing 'adequate support.' While qualitative data were comprised of think-aloud protocols and written score explanations, the quantitative data set included essay scores that were obtained analytically. With this in mind, the foci of this research are to discover whether rating experience plays a role on score variation along with the consideration of essay quality and the rating behaviors depicted by raters with different experience profiles.

As a mixed-methods research design, this study used the convergent parallel design (see Figure 1) where the level of interaction between qualitative and quantitative strands is interactive during the data collection process and the overall interpretation of the results but independent during data analysis. This design prioritizes the methods equally in terms of addressing the research problem (Cresswell, 2011, p. 541).

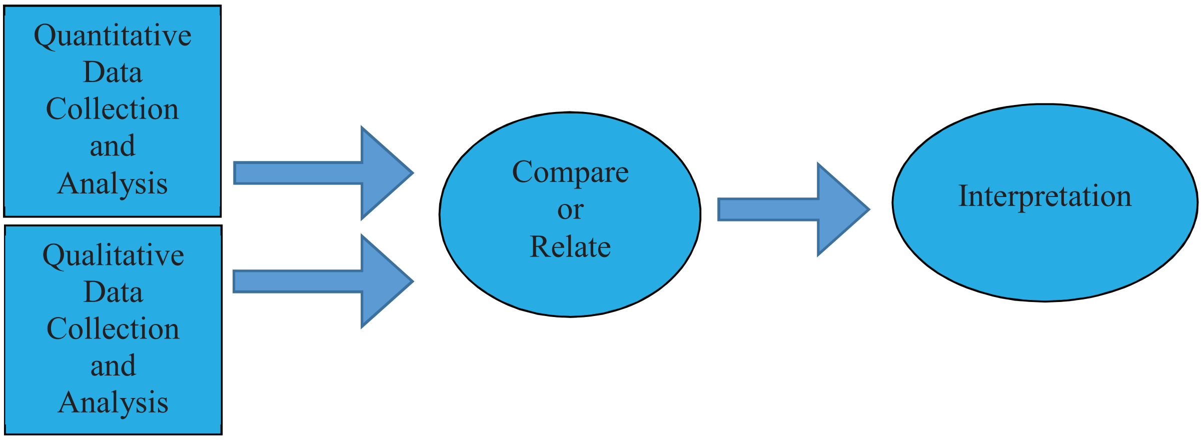

Quantitative Data Collection and Analysis
Qualitative Data Collection and Analysis
Compare or Relate
Interpretation

Figure 1. Convergent parallel design (adapted from Cresswell, 2011, p. 541).

Participant raters were selected with two aims in mind. On the one hand, the raters participating in this study were selected from a variety of universities to represent a wide context, which aims to interpret the findings from a wider perspective. On the other hand, a significant number of the participants, equaling almost half of the raters, were selected from a single university to observe the effects of institutional assessment policies on the rating process. The writing samples were collected from English Language Teaching Department of a state university and used for obtaining both qualitative and quantitative data.

This chapter begins with a section that describes the theoretical framework and the raters who participated in this study. It then continues with detailed descriptions of the instruments used for data collection. The following sections explain data collection procedures. Then, data analysis steps are explained in detail followed by the highlights of research ethics and a summary section.

Theoretical Framework

The three theoretical frameworks guiding writing assessment research are CTT approach, IRT approach, and the G- theory approach (Elorbany & Huang, 2012). In the following section, G-theory, as the theoretical framework of this study, will be explained by elaborating into its features in comparison with CTT.

Generalizability theory. While measuring a language ability, it is important to consider that what is measured is an abstract construct and it cannot be directly observed. In other words, an individual's true score for any ability cannot be directly tested (Bachman, 1990). Therefore, reliability of test scores must rely on the relationships between the *observed score* and *true score* (Bachman, 1990). While a true score occurs due to the ability of an individual and it represents the actual performance of the examinee in a measurement context, an observed score is derived from the interactions between true score and error score, which is caused by the factors other than the ability being tested (Fulcher & Davidson, 2007; Huot, 1990; Kieffer, 1998). That is to say, a true score is comprised of two variances—observed score variance and error score variance—and error score variance is described as unsystematic or random and it is not correlated with true scores (Bachman, 1990; Briesch et al., 2014; Fulcher & Davidson, 2007). Classical test theory is known to be the traditional measurement model, which assumes that an observed or actual score is equal to the combination of true score and error score as illustrated in the equation below:

$$X = T + E,$$

where X, T, and E represent observed score, true score, and error score, respectively (Brennan, 2011b; Briesch et al., 2014). There are multiple unsystematic and random sources of error score hidden in E, therefore, classical test theory is considered a weak theory as it accounts for only a single source of variance out of multiple error sources within a given analysis (Huang, 2008, 2011, 2012; Linn & Burton, 1994).

Moving from the limitations of CTT, Cronbach et al. (1972) developed G-theory. This theory functions as a theoretical framework for the test designers to assess multiple sources of variation or measurement error within a given assessment context (Briesch et al., 2014; Cronbach et al., 1972; Shavelson & Webb, 1991; Suen, 1990). In other words, superior to CTT approach, G-theory is able to identify the multiple potential sources that contribute to score variation and estimate the size of these sources of error in multifaceted measurements (Saeidi & Rashvand Semiyari, 2011; Shavelson, Baxter, & Gao, 1993).

G-theory can be seen as an extension of CTT in which only two sources of errors are concerned: "a single ability and a single source of errors" (Bachman, 1990, p. 188). However, G-theory deals with multiple sources of variance and estimates the relative contributions of these sources to the measurement simultaneously depending on the interest and specification of the test developers and test users (Bachman, 1990). To illustrate, when two or more raters score a set of essays written on two topics using holistic and analytic rubrics, the following facets can be identified as sources contributing to the variability of the scores: variability between raters, variability between scoring methods, variability between topics, and the interactions between or among these facets. It should be noted that different sources of variance such as occasion, rater, topic, and scoring method are called *facets* in G-theory and the term 'facet' is adopted in G-theory to separate the sources of errors from the factors in factor analysis (Briesch et al., 2014). Additionally, the levels of a source are considered *conditions* in that when a rater is treated as a facet, rater 1, rater 2, rater 3 etc., are identified as conditions (Güler et al., 2012).

Considering the aforementioned measurement scenario, G-theory can estimate the magnitude of the variance stemming from each facet inherently. This process is comprised of two stages: a *Generalizability study* (G-study) and a *Decision study* (D-study). G-study aims to identify and quantify the sources of variance in test scores attributed to each facet (student, test, rater, scale, etc.) in the testing environment (Barkaoui, 2007b) and it provides information for the D-study to make decisions about individuals or groups of individuals (Huang et al., 2014). In other words, G-study is used to evaluate the relative importance of various sources of measurement error and investigate the effects of diverse changes in the measurement design (e.g., different number of tasks or raters/ratings). D-study integrates the ideal design to allow the interpretation of score reliability in the norm-referenced or criterion-referenced frame of reference (Brennan, 2001b; Briesch et al., 2014; Gao & Brennan, 2001; Huang, 2008).

In addition to G-studies and D-studies, the other important considerations in the G-theory framework are the concepts of *universe of admissible observations* and *universe of generalization* (Brennan, 2000, 2011). While the former term refers to the range of conditions under which a certain construct may be measured, the latter can be explained as the conditions of a facet to which a decision-maker desires to generalize (Brennan, 2000; Briesch et al., 2014; Shavelson & Webb, 1991). Although CTT attempts to estimate a *true score*, G-theory focuses on the *universe score* that is expected from the objects of measurements—examinees or students—across all admissible measurement procedures (Briesch et al., 2014; Shavelson & Webb, 1991). Even though *true score* or *universe score* is considered to be the ideal score that should be assigned to the test-taker (Huot, 1990), it is not very likely for an observed score to match with the universe score perfectly (Briesch et al., 2014; Greenberg, 1992, Huang, 2009; Huang & Foote, 2010). In other words, while generalizing an observed performance to universe score, some degree of error is likely to occur, the extent of which can be calculated through *generalizability coefficients* and *dependability coefficients* (Briesch et al., 2014; Huang & Foote, 2010; Shavelson & Webb, 1991). *Generalizability coefficients* are used in a norm-referenced test in which the scores of each test-taker are interpreted relative to the other test-takers' performance, whereas *dependability coefficients* are used in a criterion-referenced test context in which each test-taker's score is interpreted relative to a fixed set of predetermined test criteria (J. D. Brown, 1996; H. D. Brown, 2004; Shavelson & Webb, 1991).

The analysis in G-theory can be designed with the facets fully *crossed* or *nested*, which can be explained with the interaction type of the conditions of facets within the given design. If every condition of a facet interacts with the conditions of other facet(s), the design is then fully crossed, whereas only some conditions in a facet are observed with only some conditions of other facet(s) in nested designs (Briesch et al., 2014; Güler et al., 2012; Kieffer, 1998). To illustrate, assuming that all the students (*s*) wrote essays on two topics (*t*) and all the raters (*r*) scored all essays that every student wrote on both topics, the design would be crossed as student-by-rater-by-topic (*s x t x r*). Conversely, a nested design occurs when different students (*s*) write essays on different topics (*t*) and different raters (*r*) assess the essays written by different students (*s : t : r*). Additionally, there may be mixed designs in G-studies (Güler et al., 2012). Another distinction that can be made in the G-theory framework is between the facets in that they can be regarded as *fixed* or *random* (Briesch et al., 2014; Güler et al., 2012). If the researcher is dealing with the instances under investigation and does not desire to generalize beyond those instances, then the facet is treated as fixed while all conditions in a facet are exchangeable with the ones in the universe when the facet is considered random (Briesch et al., 2014; Güler et al., 2012).

In conclusion, G-theory is an appropriate approach for the context of this study considering the aforementioned features of the framework. It enabled the researcher to detect the sources of variance and their magnitudes to the variability of scores under investigation. It also allowed the researcher to optimize the best measurement conditions within given facets in the study.

Selection of Raters

In this research, convenience sampling was used given the proximity and availability of the setting and volunteer participants to the researcher. As Dörnyei (2007) suggests, convenience sampling often includes elements of purposive sampling: “besides the relative ease of accessibility, participants also have to possess certain key characteristics that are related to the purpose of the investigation” (p. 99). Such was the case in this research, where participants were full-time employees at the university level with varying degrees of rating experience. The researcher ensured privacy and confidentiality, which were central to the ethics of research practice to protect the participants' identities. All names used in this study were pseudonyms. The study included a total number of 34 participants initially; however, one participant moved abroad for his Ph.D. education and had to drop out of the study. The remaining 33 participants were working at Bursa Technical University (BTU) and other higher education institutions located in different regions of Turkey. Figure 2 illustrates the locations of the universities at which the participants were employed.

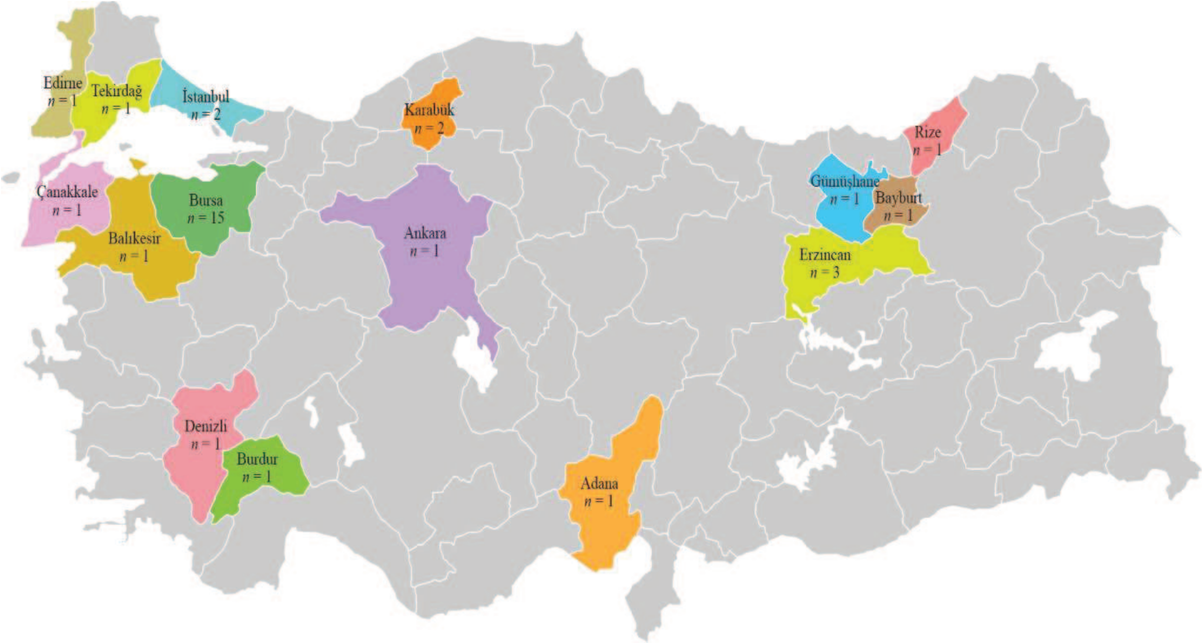


Figure 2. Locations of universities at which the participants are employed (adapted from blank map of Republic of Turkey's provinces, by Baydin, 2006)

The participants were based in 16 different state universities in 15 different cities. While each city represented only one university, the two participants from the city of Istanbul worked at two different universities during the present research study. All the raters who participated in this study were professionals in the field of interdisciplinary English language teaching, learning and assessment, and regular employees at the School of Foreign Languages (SFL), Foreign Languages (FL) Department, or ELT Department at a state university in Turkey. These 33 raters were all graduates from different ELT and ELL departments in

Turkey, and they have the same L1 background (Turkish). The participants varied in their professional experience in teaching and assessing EFL writing. The raters were provided with a rater profile form (Appendix A) adapted from previous studies (Barkaoui, 2008; Cumming et al., 2001) in order to collect data about their backgrounds including personal, educational, and professional information in general. Additionally, they were asked to indicate their experience assessing EFL writing in years as well as their perceptions of themselves as EFL writing assessors.

An analysis of the participants' experience grading papers revealed three experience levels, which were then used to group raters into low-, medium-, and high-experienced groups. The breakdown of rater experience is presented in Table 2.

Table 2

Rating Experience of the Participants

		Experience rating EFL essays					
		2 years or less	*3-4 years*	*5-6 years*	*7-10 years*	*10+ years*	***Total***
Experience Group	*Low*	**3**	**10**	0	0	0	**13**
	Medium	0	0	**10**	0	0	**10**
	High	0	0	0	**6**	**4**	**10**
Total		3	10	10	6	4	**33**

As can be seen in Table 2, of the 33 participants, 13 reported four years or less experience grading papers for EFL writing assessment. These raters were categorized as low-experienced raters. Ten raters reported 5-6 years of experience rating papers for EFL writing assessment, and the remaining ten raters reported seven years or more rating experience. These groups were categorized as medium- and high-experienced raters, respectively.

Table 3 illustrates the distribution of raters' gender and age based on the experience group.

Table 3

Gender and Age Distribution of the Participants

		Gender		Age		
		Male	*Female*	*20-30 years*	*31-40 years*	*41-50 years*
Experience Group	*Low*	6	7	9	3	1
	Medium	6	4	6	4	0
	High	7	3	1	8	1
Total		**19**	**14**	**16**	**15**	**2**

A number of 19 male and 14 female raters participated in this research. There were six male and seven female raters in the low-experienced group, while the medium-level experienced group included six males and four females. The gender distribution came out to be seven males and three females for high-experienced rater group. When considering raters' ages, 16 raters were between 20 and 30 years old; the remainder 17 raters were over 30 years old. Nine low-experienced raters were between 20-30 years old while three raters ranged between 31-40 years old; only one rater was over 41 years old. As for the medium-experienced group, six raters were between 20-30 years of age and four raters were between 30-40 years old. Among the high-experienced raters, there was only one rater between the 20-30 and within the 41-50 year ranges each, while the remainder eight raters in this group were between 31-40 years old. Furthermore, Table 4 shows the distribution of academic title and institutional affiliation of the raters based on their experience.

Table 4

Participants' Academic Title and Institutional Affiliation Distribution

		Academic Title		Institution	
		Research assistant	*Instructor*	*BTU*	*Other*
Experience Group	*Low*	2	11	5	8
	Medium	0	10	4	6
	High	0	10	6	4
Total		**2**	**31**	**15**	**18**

According to Table 4, while the great majority of the participants (*n* = 31) were working as EFL instructors at tertiary level education in Turkey, two of the participants were employed as research assistants at their universities. While the two research assistants reported themselves as low-experienced raters, 11 of the instructors were grouped in the low-experienced category and the remaining 20 instructors fell into medium- and high-experienced groups equally. When considered their institutional affiliation, five raters from the low-experienced group, four from the medium-experienced group, and six from the high-experienced group were working at BTU while the remaining eight raters from the low-experienced group, six from the medium- experienced group, and four from the high-experienced group were from different higher education institutions. To highlight the raters' educational background, Table 5 illustrates the participants' highest level of degree obtained and their rater training history.

Table 5

Participants' Highest Level of Education and Previous Training on Writing Assessment

		Degree		**Previous Training**	
		BA	*MA*	*Yes*	*No*
Experience Group	*Low*	6	7	7	6
	Medium	5	5	6	4
	High	3	7	6	4
Total		**14**	**19**	**19**	**14**

When considering the highest level of education that the raters completed, Table 5 shows that 14 of the raters held a BA degree while 19 of the raters were MA graduates. Additionally, 19 raters reported that they received training on writing assessment while 14 raters had not received training on assessing writing previously. Moreover, Table 6 illustrates detailed information about the participants' previous experiences in terms of EFL and writing teaching separately.

Table 6

Teaching Experience of the Participants

		Teaching EFL (total)				
		2 years or less	*3-4 years*	*5-6 years*	*7-10 years*	*10+ years*
Experience Group	*Low*	2	4	0	3	4
	Medium	0	0	5	4	1
	High	0	0	0	3	7
Total		**2**	**4**	**5**	**10**	**12**
		Teaching EFL in University Settings				
		2 years or less	*3-4 years*	*5-6 years*	*7-10 years*	*10+ years*
Experience Group	*Low*	3	4	0	4	2
	Medium	0	0	10	0	0
	High	2	2	0	5	1
Total		**5**	**6**	**10**	**9**	**3**
		Teaching Writing (total)				
		2 years or less	*3-4 years*	*5-6 years*	*7-10 years*	*10+ years*
Experience Group	*Low*	5	8	0	0	0
	Medium	1	3	6	0	0
	High	2	2	0	4	2
Total		**8**	**13**	**6**	**4**	**2**
		Teaching Writing in University Settings				
		2 years or less	*3-4 years*	*5-6 years*	*7-10 years*	*10+ years*
Experience Group	*Low*	6	7	0	0	0
	Medium	1	3	6	0	0
	High	3	2	0	4	1
Total		**10**	**12**	**6**	**4**	**1**

According to Table 6, the total experience of the raters in teaching EFL and EFL writing varied both in general and in university contexts in that 22 of the raters (66.7%) had over seven years of teaching experience in general while 21 raters (63.7%) had less than seven years of experience in teaching EFL at the university level. As for experience in teaching writing, 21 of the raters had less than five years' experience while 12 raters had over five years' experience teaching writing in general. Furthermore, while 22 raters had less than five years' experience at the university level, 11 raters had over five years' experience in teaching writing at the university level.

In addition, the participants were asked to describe their experience as raters on a 5-point Likert-type scale ranging from "No experience" to "Very experienced." In this way, the researcher was able to obtain a measure of the participants' self-perception of their experience as raters. Their responses are presented in Table 7 with respect to their experience grouping.

Table 7

Participants' Self-described Rating Experience

		Self-described Experience				
		No experience	*Little experience*	*Some experience*	*Experienced*	*Very experienced*
Experience Group	*Low*	2	3	4	4	0
	Medium	0	1	5	4	0
	High	1	0	5	3	1
Total		**3**	**4**	**14**	**11**	**1**

Interestingly, there was only moderate overlap between the raters' reported experience in rating papers and their self-perceptions of their experience as raters. In total, 11 raters described themselves as "experienced" raters, and an additional rater described him or herself as a "very experienced" rater. However, of the 11 raters who described themselves as experienced raters, only three raters belonged to the high-experienced rater group. Four of the self-described high-experienced raters belonged to the low-experienced group and the remaining four belonged to the medium-experienced group. Fourteen raters described themselves as having some experience rating papers. These 14 raters were distributed across the experience groups, with four raters belonging to the low-experienced group, five raters belonging to the medium-experienced group, and five raters belonging to the high-experienced group. Of the seven raters who described themselves as having little or no experience rating papers, five belonged to the low-experienced group, one belonged to the medium-experienced group, and one belonged to the high-experienced group. This final rater presents an interesting case, as he or she reported at least seven years of rating experience and yet perceives him or herself as having no experience rating papers. Given the distribution of raters according to their self-described experience, three categories of self-described experience were created as in Table 8.

Table 8

Categories of Self-described Rater Experience

		Self-described Experience					
		No experience	*Little experience*	*Some experience*	*Experienced*	*Very experienced*	*Total*
Self-described Group	*Low*	3	4	0	0	0	7
	Medium	0	0	14	0	0	14
	High	0	0	0	11	1	12
Total		**3**	**4**	**14**	**11**	**1**	**33**

The category of self-described low-experienced raters (*n* = 7) included raters who described themselves as having no or little experience rating papers; the category of self-described medium-experienced raters (*n* = 14) included raters who described themselves as having some experience; and the category of self-described high-experienced raters (*n* = 12) included the participants who described themselves as experienced or very experienced raters.

The background information pertaining to the participants were provided to elaborate into the demographic characteristics of the raters. However, only some of the aforementioned characteristics including reported and self-described rating experience were used in the analysis and interpretation of data collected quantitatively and qualitatively.

Data Collection Instruments

Adopting a mixed methodology, this study collected both qualitative and quantitative data using a background questionnaire, analytic scores given to EFL essays, TAPS, and written score explanations. Firstly, the quantitative data consisted of 50 essay scores assigned to EFL essays by each rater using the adapted version of 10-point analytic scoring scale developed by Han (2013) which was used to collect data from raters to be analyzed within the G-theory framework (see Appendix B). Secondly, the raters were requested to provide information about their background using the rater's profile form immediately after they completed the rating task. This data set was used to divide the raters into groups based on their characteristics to interpret qualitative and quantitative data. As for the qualitative data, TAPs were used to investigate rater cognition and their decision-making behaviors. Raters were also asked to justify their scores by listing three written explanations (reasons) for their scores assigned to the essays and this data set was used to supplement the data obtained from TAPs. Detailed information about the data collection instruments is provided in the following sections.

Selection of EFL essays. The writing samples were collected from EFL students enrolled in the Advanced Reading and Writing Skills Course taught at Çanakkale Onsekiz Mart University Faculty of Education Foreign Languages Teaching Department, English Language Teaching Program. In the literature, essay topic is listed as one of three general sources for score variability of writing tests along with raters and students, and different writing topics may affect learners' writing skills, resulting in different writing scores (McColly, 1970). Therefore, a good writing topic should allow the students to show their performance at the maximum level (Weigle, 2002). As such, the students were provided with a topic that could arouse their professional and educational interests as in the following:

> *Some people think that English teachers working at primary schools and high schools are insufficient to teach English effectively. Therefore, Ministry of Education in Turkey is thinking of hiring native English-speaking teachers to support English language education. Do you think that English teachers in Turkey are qualified enough for teaching English to the students or should English language education in Turkey be supported by native English-speaking teachers? Use specific reasons and examples to develop your essay.*

Additionally, before the students were tasked to write their essays, the researcher prepared a platform in the classroom environment to allow the students to exchange their ideas on the essay topic. In doing so, the researcher grounded this procedure in Vygotsky's Zone of

Proximal Development where the social interaction is essential for constructing knowledge (Vygotsky, 1978) so that the students helped each other establish the required background knowledge to fulfill the writing task about the given topic above. Added to that, it is believed that the most substantial problem with independent writing tasks is topic familiarity, which can cause students to demonstrate a poor performance (Gebril, 2009). Therefore, source-based writing can be considered effective on students' building background knowledge (Weigle, 2004), resulting in achieving fair judgements in more equal testing conditions (Plakans, 2007). The essays were not written within a limited time in the classroom; the students were given a 3-day period to write the essays on their computers instead. This procedure eliminated the possibility that raters' scoring behavior might be affected by students' handwriting (Song & Caruso, 1996). The essays were 500- to 700-word length compositions and accepted through a text-matching software—*Turnitin*—to ensure the originality of the essays. At the end of the submission process, a total number of 104 essays were gathered from the students to be used in the study.

Because one of the aims of this research is to see the impact of distinctive essay quality on the rating process including the variation between scores and varying rater behaviors towards essays of different qualities, the collected essays went through a division process carried out by three independent quality-check raters, two of whom held a PhD degree in the Department of ELT and one had a PhD degree in Applied Linguistics. These raters were professionals in the field of interdisciplinary English language teaching, learning and assessment, and regular employees of an SFL, ELT Department, and ELL Department at different state universities in Turkey and had over 10-year expertise in teaching and assessing ESL/EFL writing. The quality-check raters were provided with a set of assessment instructions (Appendix C) and a holistic scoring scale (Appendix D) along with the essay pack. This scale was developed by the BTU writing team (BTU SFL, 2014) for grading large-scale tests such as entrance, final, and exit exams, which do not require giving feedback to the authors. It was comprised of language and topic development sections with a 3-point weight each. Relying on the scale, the expert raters were expected to divide the essays into three quality groups—high, medium, and low—but not to score them.

Of the 104 essays collected from the students, 50 essays, 25 of which were low and the other 25 of which were high, were selected to be used in this research. The essays were accepted in two ways: 1) all the raters assigned the same quality categorization as high or low, or 2) when 2/3 of the raters agreed on high- or low-quality, the essay was sent to the 4th independent rater and the 4th rater confirmed the decision of the 2/3 majority. Otherwise, the essays were rejected and discarded from the study. After the expert raters completed the classification process, the researcher excluded medium quality (n = 28) essays from the study. Twenty-three essays were also left out of the study given that there was a discrepancy among the raters in their decisions (e.g., three raters reported three different qualities such as low, medium, and high). Initially, 3/3 of the expert raters assigned high-quality ratings to 10 essays and low-quality ratings to 14 essays, and these essays were determined to be used in the study. However, 2/3 of the raters assigned high-quality ratings to 18 essays and low-quality ratings to 11 essays, requiring a 4th independent assessor to make the final decision about the essays. As a native-speaker of English, the 4th rater was pursuing her Ph.D. studies in the Department of Education with a primary research concentration of Applied Linguistics and had four years of EFL teaching experience at the tertiary level in Turkey. While she confirmed the 2/3-majority decision for all of the low-quality essays, she disagreed with the 2/3-majority decision for three high-quality essays and confirmed the majority decision for 15 essays out of 18.

Overall, the quality-check process before collecting main data was carried out carefully with the involvement of four expert raters so that the researcher made sure to address essays divided in their qualities in a meticulous way in order to serve the purpose of the study.

Twenty-five low-quality and 25 high-quality essays were chosen out of 104 essays to be used for the main data collection. Figure 3 summarizes how the essays were categorized based on their qualities.

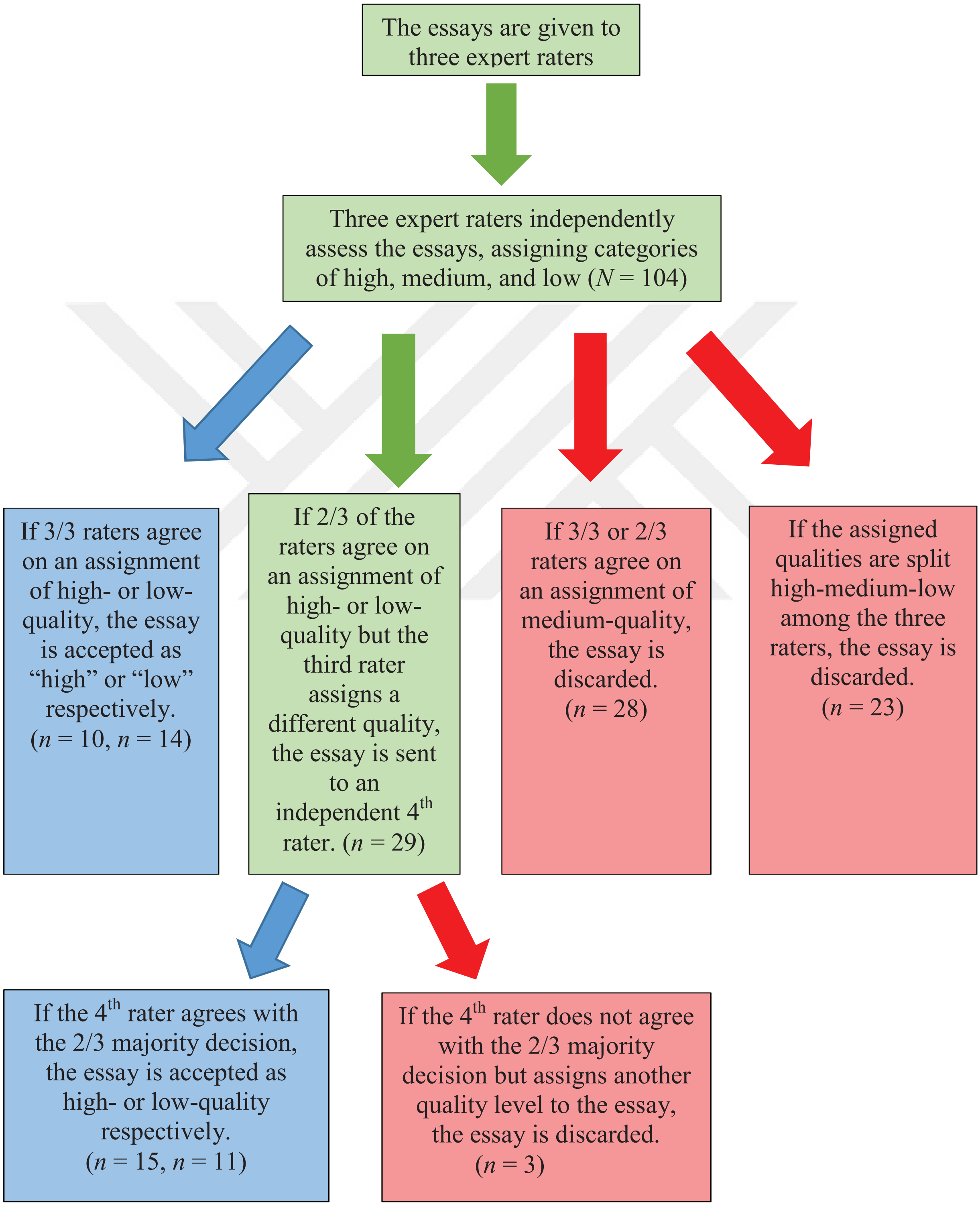

The essays are given to three expert raters
Three expert raters independently assess the essays, assigning categories of high, medium, and low (N = 104)
If 3/3 raters agree on an assignment of high- or low-quality, the essay is accepted as "high" or "low" respectively. (n = 10, n = 14)
If 2/3 of the raters agree on an assignment of high- or low-quality but the third rater assigns a different quality, the essay is sent to an independent 4th rater. (n = 29)
If 3/3 or 2/3 raters agree on an assignment of medium-quality, the essay is discarded. (n = 28)
If the assigned qualities are split high-medium-low among the three raters, the essay is discarded. (n = 23)
If the 4th rater agrees with the 2/3 majority decision, the essay is accepted as high- or low-quality respectively. (n = 15, n = 11)
If the 4th rater does not agree with the 2/3 majority decision but assigns another quality level to the essay, the essay is discarded. (n = 3)

Figure 3. Quality classification of EFL essays.

Rating scale. A 10-point analytic scoring scale (Appendix B) adapted from Han (2013) was used in this study because analytic rubrics are considered more suitable than holistic criteria to “assess accurately the quality of L2 writing for purposes such as research, high-stakes testing or diagnostic assessment, where the quality of information from evaluation is more important” (Shi, 2001, p. 317). Han modified the instrument benefiting from the rubric development literature, course objectives, sample EFL essays written by Turkish students, and contributions of department members in his research context. Originally, the rubric consisted of five scoring criteria with different maximum point distributions: grammar (3 pts.), content (2 pts.), organization (2 pts.), style and quality of expression (1.5 pts.), and mechanics (1.5 pts.). Each component had five scoring bands with varying cut points and score intervals (e.g. 0 – 0.4, 1.2 – 1.7, 2.5 – 3.0). While carrying out the rubric adaptation process, the researcher had two main purposes in mind: 1) adapting the rubric with the involvement of the participating raters and 2) orienting the raters to the rubric prior to the main data collection. All the participants of this study ($N = 33$) were included in the adaptation process to ensure the validity and reliability of the tool in that the raters were expected to use a rubric with which they would feel comfortable rather than base their scores on a rubric of which they would be critical and unfamiliar (Barkaoui, 2007b; Davidson, 1991; Hamp-Lyons, 1991; Weigle, 2002). In doing so, the researcher used three essays of distinct qualities, the rubric orientation instructions list (see Appendix E), and the original version of Han’s rubric. The only change that was made on the tool prior to sharing it with the participating raters was to arrange an equal score distribution to the components (maximum 2 pts. for each component), which aimed to eliminate any potential rater biases that would stem from weight distribution of the rubric. Figure 4 illustrates the rubric orientation process that was designed by the researcher to form the final draft of the rubric.

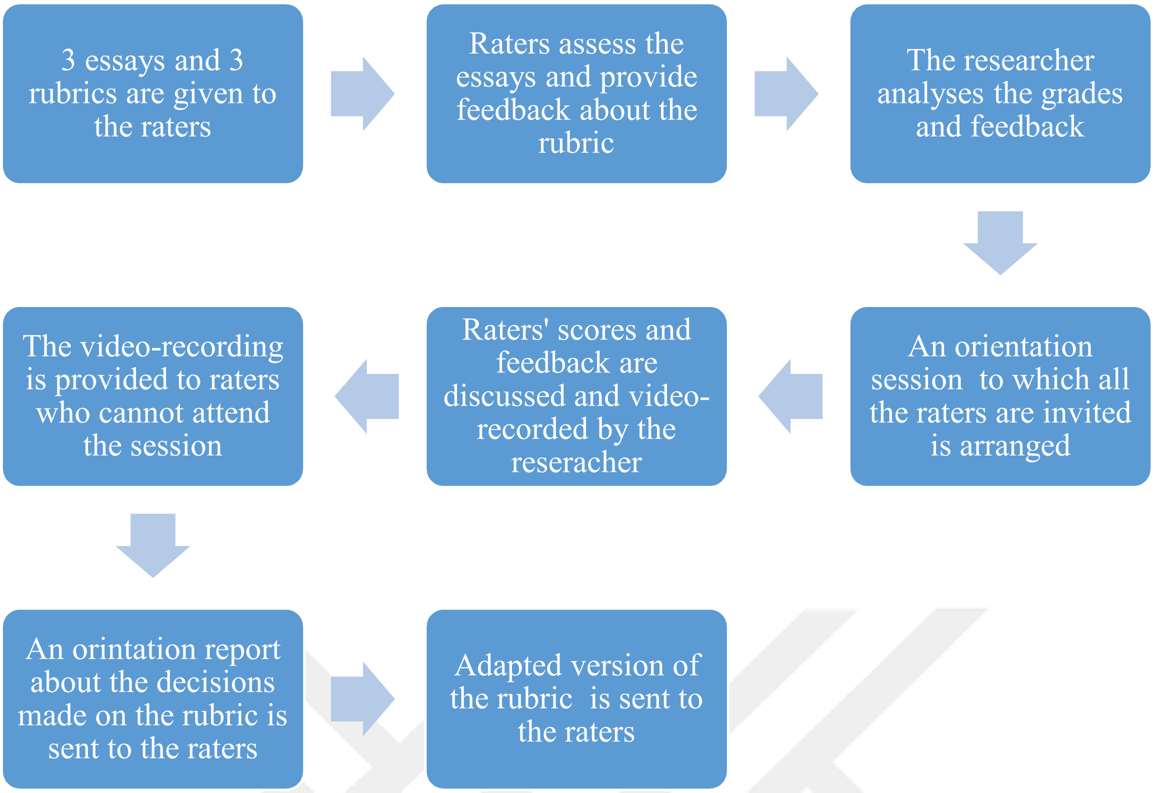

3 essays and 3 rubrics are given to the raters
Raters assess the essays and provide feedback about the rubric
The researcher analyses the grades and feedback
An orientation session to which all the raters are invited is arranged
Raters' scores and feedback are discussed and video-recorded by the reseracher
The video-recording is provided to raters who cannot attend the session
An orintation report about the decisions made on the rubric is sent to the raters
Adapted version of the rubric is sent to the raters

Figure 4. Rubric orientation process.

In order to see the practicality of the rubric, the raters were given three essays of varying qualities, and they were asked to evaluate the essays using the rubric. Furthermore, the raters indicated three written explanations for their scores with regards to positive and negative aspects of the essays. Additionally, they were provided with a rubric feedback form (see Appendix F) to reflect their opinions and comments on the rubric so that the researcher could make necessary modifications to the tool. Based on the feedback that the raters provided and their responses to the essays, the researcher organized a face-to-face rubric orientation session and the session was video-recorded and uploaded to *YouTube* (Şahan, 2016a) to be shared with the raters who were living in different cities and could not attend the session. In this way, the researcher aimed to open raters’ suggestions about the rubric for discussion and put the final touches to the rubric democratically.

The evaluation of three essays allowed the raters to assess the weaknesses and strengths of the rubric. Based on their scoring practices with the rubric, the raters provided feedback about the rubric including the practicality of the tool, clarity of the descriptors, and weight distribution of the rubric components. When the essay scores given by 33 raters were analyzed, a large difference was found between the maximum and minimum score for each paper. Table 9 summarizes statistical analysis of the scores.

Table 9

Descriptive Statistics for Assigned Scores to Rubric Orientation Essays

Essay	N_{Rater}	*Min.*	*Max.*	*M*	*SD*
Essay 1 (low-quality)	33	2.20	8.60	5.31	1.35
Essay 2 (medium-quality	33	2.20	8.00	5.34	1.39
Essay 3 (high-quality)	33	5.20	9.80	7.82	1.19

When the scores given to the essays were analyzed, the mean scores for Essay 1 and 2 were found to be similar while Essay 3 had a higher mean score. Additionally, the gaps between minimum and maximum scores assigned to Essay 1 and 2 were larger than that of Essay 3, indicating smaller score variation for Essay 3. These findings showed that raters varied less with their scores assigned to the high-quality essay. When asked how practical the rubric was, the raters gave different responses. Figure 5 illustrates the ratings of the participants for the practicality of the rubric.

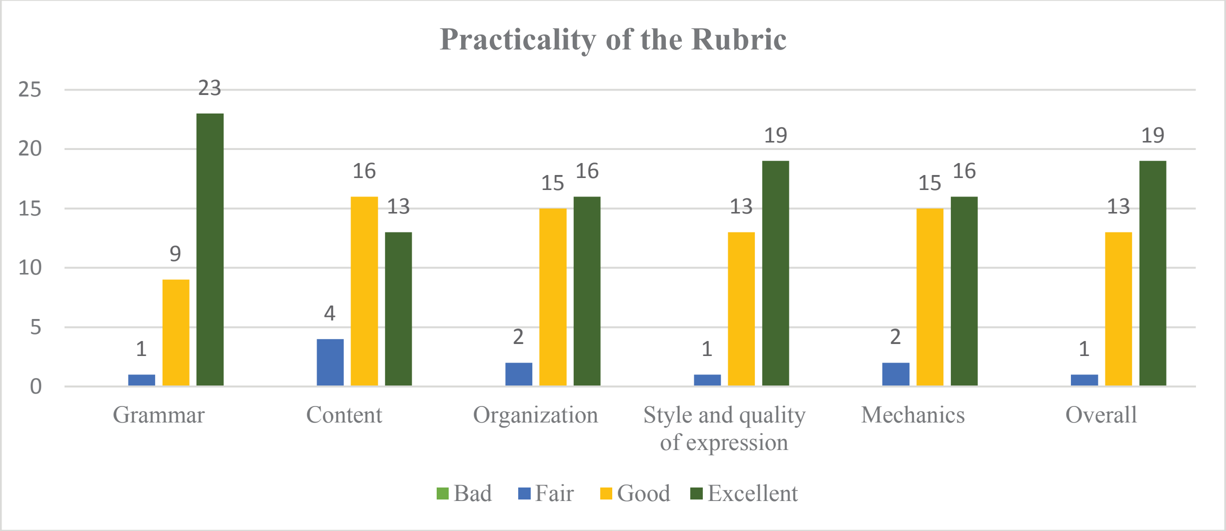


Figure 5. Ratings for the practicality of the rubric.

As can be seen in Figure 5, the rubric was generally considered good or excellent by most of the raters. Additionally, the responses given by the raters for each component of the rubric displayed similar tendencies toward the practicality of the specific aspects of the tool. However, the ratings indicated that the scoring scale needed some improvements and modifications to fit the research context of this study in that the raters provided the researcher with informative explanations on how to improve the tool. Although 84% of the raters (*n* = 28) thought that the expressions for each performance level in the rubric were distinctive enough to identify the strengths and weaknesses of the essays and helped them make their decisions about the essays, they commented on the wording of the descriptors and offered better word choices in order to be clearer and more specific in the descriptors. Furthermore, 65% of the raters suggested that the weight distribution should be changed and Figure 6 displays the new score weight distribution of the five components of the analytic rubric, organized based on the participants' feedback.

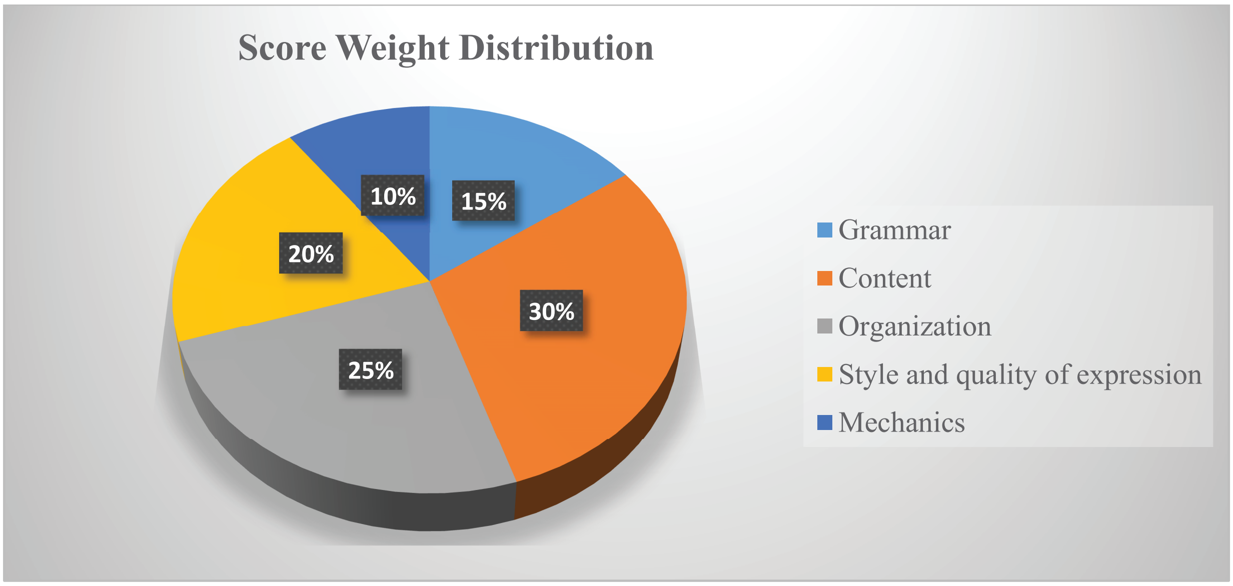


Figure 6. Distribution of score weights to the rubric's subscales.

The problems that each rater underlined on the feedback form were argued and pre-proposed solutions and/or immediate proposals about the rubric were discussed. Based on the discussion carried out in the session, the researcher prepared a rubric orientation and adaptation report (Appendix G) and provided it to the raters three days later. In this way, all the raters were able to see the changes made on the tool in a clear way.

Following the completion of each participant's assessments of the 50 essays, the researcher contacted the raters to inquire into the practicality of the rating scale after it had been used to assess the 50 papers of different qualities. Using a Likert-type scale with anchors arranged from least important to most important (1 = never, 2 = rarely, 3 = partially, 4 = mostly, and 5 = always), the raters were asked to indicate to what extent they agreed that a) the rubric served the purpose of the assessment task and b) they felt secure during the assessment task while using the rubric. In this phase, 31 out of 33 raters responded to the researcher's email. The ratings indicated that the scoring scale used in this study was functional with a mean value of 4.51 and the raters felt secure while using the scale with a mean value of 4.41 out of 5.

In addition, the raters reported on the reliability of the scoring rubric with respect to its effectiveness in helping them assign scores that they believed students' essays actually deserved. The comments revealed that all the raters responded to this particular post-scoring inquiry were favorable towards the scoring rubric as illustrated as follows:

> I think that various aspects that the rubric cover and well-arranged score ranges contributed to less margin of error in my scorings. As such, I am planning to use this rubric to assess my own students' written products as well. (Kamil)
>
> The rubric consisted of five main evaluation criteria and instead of assigning a holistic rating, it allowed me to give cut scores for the five components. In this way, I was able to assess and rate the essays based on the given criteria and the clear descriptors. Although I do not have any previous experience in assessing writing, I did not have any trouble rating the essays given that the rubric was comprehensible and practical. (Adalya)
>
> I generally tend to assign higher scores in subjective performance assessments like writing. When an essay is double-scored by another rater, the score that I assign is always considerably higher than that of the other rater. However, for the first time I felt that I assigned scores that the essays deserved. (Ozge)
>
> Writing assessment is considered subjective but this rubric helped me assign fair scorings. I think that if I had had such rubric in my previous assessment tasks, I would have assigned more standard and fair scoring to the students' essays. (Efe)

Apart from the quoted rater comments, the remainder 27 responses were also positive about the scoring scale and focused on several issues. The raters thought that the rubric was user-friendly and practical, clear and comprehensible, systematic and detailed, and objective and distinctive. Furthermore, three raters indicated that it was the best scoring scale they had ever used. The ratings and the comments that appeared in the post-scoring inquiry show that involving the raters in the process of developing and/or adapting the scoring criteria might increase their trust in and reliance on the scoring scale while doing their evaluations.

In brief, the rubric adaptation process was performed collaboratively by following pre-determined steps in which the raters were introduced to the aim of the assessment purpose and task, the students' proficiency levels, and the course objectives and outcomes before they scored three sample essays selected from the essay pool in this research context. Following that, a discussion was carried out, which allowed each rater to give voice to their opinions on the rubric and every one of the criticisms was shared with the other participant raters. These steps helped the raters get oriented to the rubric before the main data collection phase commenced. In addition, the reactions of the raters to the post-rating inquiry about the scoring scale indicated that the rubric served the purpose of the assessment task in this particular research context.

Think-aloud protocols. Charney (1984) argues that quick and superficial rating is essential to arrive at reliable scores instead of developing deeper consideration of the text. However, other researchers suggest that reliable and valid scores can be obtained only when raters base their judgements on rich interpretations of the texts as well as by using scoring rubrics (Cumming et al., 2001; Huot, 1993). In this sense, TAPs can be considered important in order to understand the rating process to address some of the assessment concerns related to scoring. Similarly, Connor-Linton (1995) underscores the importance of understanding what raters are doing during their assessments to make sense of their scores. In this regard, the thinking-aloud process is a kind of cognitive task that is comprised of several mental states, each of which is the end product of processed information (Wolfe et al., 1998). This qualitative method has been used in writing assessment studies in which raters verbalize their thoughts during their assessments and their verbalizations are recorded simultaneously. Following that, the recordings are transcribed in meaningful units and coded according to a scheme developed previously or within the preliminary findings of the study (Weigle, 1994).

Think-aloud protocols have been increasingly used in both first (e.g., Huot, 1993; Wolfe, et al., 1998) and second (e.g., Barkaoui, 2007b, 2010a, 2010b; Connor-Linton,1995; Cumming, 1990; Cumming et al., 2001, 2002; Gebril & Plakans, 2014; Han, 2017; Lumley, 2005) language writing rating processes. In this type of data collection, raters receive instructions (see Appendix H) to verbalize their thoughts while completing the task of rating a set of essays in the context of this study. Raters' spoken thoughts are recorded, transcribed, and then analyzed to identify the decision-making processes that raters go through and the aspects of writing they attend to when rating essays (Barkaoui, 2011b).

An important concern of the researcher while designing the methodology of the current study was whether the use of TAPs would succeed in revealing the raters' cognitive map in terms of the assessment strategies that they used in their EFL writing evaluations. In this sense, the researcher planned a meticulous process in order to enable the raters to grasp the idea of thinking aloud fully and its procedures for this study in particular. Figure 7 summarizes the TAP training process.

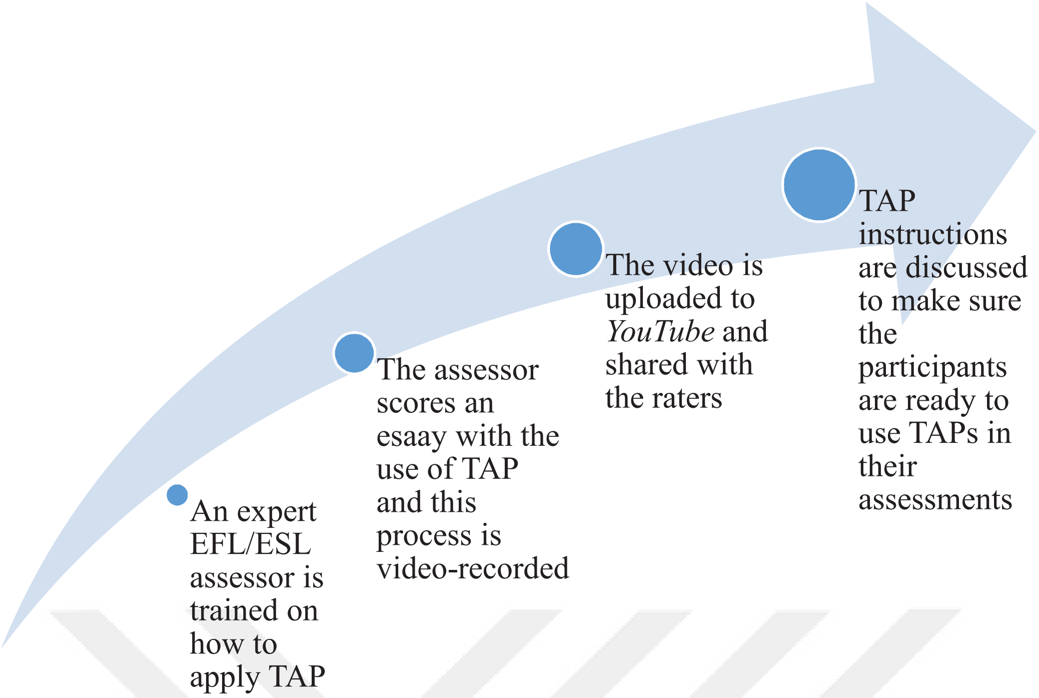

An expert EFL/ESL assessor is trained on how to apply TAP
The assessor scores an esaay with the use of TAP and this process is video-recorded
The video is uploaded to *YouTube* and shared with the raters
TAP instructions are discussed to make sure the participants are ready to use TAPs in their assessments

Figure 7. Think-aloud protocol training.

Before collecting the main data from the participants, the researcher organized a training session on how to conduct a TAP following the analytic rubric adaptation and orientation phase. In the first step, the researcher filmed a sample TAP carried out by an EFL/ESL instructor who had over four years of experience in teaching and assessing EFL/ESL writing. Firstly, the rater was introduced to the purpose of the TAP in the research context and provided with the set of instructions that should be regarded during the assessment. Secondly, a camera was set in the room to video-record the assessment task and the rater was provided with a student's paper selected from the essay pool of the study. Finally, the researcher left the rater alone in order to make him feel comfortable while grading the essay. Thereafter, the researcher uploaded the video to *YouTube* (Şahan, 2016b) for the participants to have an idea on the use of TAPs while assessing an essay. In the second step, the researcher organized a one-to-one meeting with the raters working at BTU to discuss the sample TAP video and the instructions guiding the raters on how to conduct TAPs. The raters who were participating in the study from other universities were contacted through video- or voice-calls to help them understand how to assess EFL essays while using TAPs.

Data Collection Procedures

After collecting the original EFL essays from the students and completing the quality division of the essays, the researcher prepared data packs, which were comprised of 50 essays, 50 analytic scoring rubrics, a background information questionnaire, an assessment instructions list, and a TAPs instructions list. The raters were also provided with voice-recorders if they did not have a device with voice-recording features. While the packs were handed to 15 raters at BTU in person, the packs for 18 raters working at 15 different universities located in 14 different cities were shipped in the middle of July 2016. The researcher gave a two-month period of time between the mid-July and mid-September 2016 to the participants to complete the scoring the essays. The data collection process was specifically scheduled during this period so that the raters would have a flexible and stress-free time during which to supply their data, because they were full-time employees at their institutions and it was assumed that they had less teaching and assessment responsibilities at their schools during this given time in the summer.

The data for this study were collected using quantitative and qualitative methods. The quantitative data set included a total number of 9,900 scores (1,650 total scores and 8,250 sub-scores) while the qualitative data consisted of 446 TAPs which were voice-recorded during the assessment of the essays, and 5,425 written score explanations for the assigned scores. Despite the instruction video and the detailed guidelines provided to the raters, five of them failed to record their thoughts in the way that the research required. While two of these raters recorded only one audio listing their reasons for their assigned scores to the essays, three raters commented on the essays following the completion of the assessments. Moreover, two raters different from the aforementioned five raters failed to conduct a TAP for one of the essays in the required format in that they reflected on their assessment instead of commenting during their scoring process. In this regard, a total number of 82 recordings (15.6% of the whole TAPs) were left out of the analysis. In terms of language use while recording their TAPs, five raters verbalized their thoughts entirely in English while the remaining 23 raters used translanguaging practices to assess the papers.

Each one of the raters used TAPs while assessing the pre-determined 16 essays in their essay packs. Moreover, the raters were asked to provide three reasons that affected their decisions on the essays most in order to triangulate the data obtained from TAPs. The researcher contacted the participants at certain time intervals to manage any problems that may arise from the TAPs and to address wrap-up trainings on the use of TAPs when necessary. The following two sections give further details about the data collection procedures.

Rating procedure. Using a 10-point analytic scoring scale, the raters assigned their scores to a number of 50 EFL essays by considering five different aspects including grammar, content, organization, style and quality of expression, and mechanics (1,650 total scores and 8,250 sub-scores). The essays that were of two distinct qualities—high and low—were bundled randomly in order to avoid any biases that could stem from arrangements of the essay in the sets. The raters were allowed to use partial points in assigning their scores for each component within the given score bands in the rubric (e.g. grammar = 1.2, content = 2.0, organization = 1.8, style and quality of expression = 0.9, mechanics = 0.7, and total score = 6.6). The raters were required to follow the assessment instructions (Appendix J) while scoring the essays so that the evaluation process aimed to be conducted in a standardized manner specified by the researcher. Within these instructions, the raters were informed about students' language proficiency level, their department, the essay topic, and how the essays were submitted to the researcher.

The raters were instructed to depend on the rubric separately for each time that they assessed the essays given that evaluating essays in certain qualities may impact the following assessments, resulting in unfair judgement (Daly & Dickson-Markman, 1982; Freedman, 1981; Hughes & Keeling, 1984). In the same vein, contacting the other participants to negotiate on the essays was not allowed because it might hinder the raters from relying on their own ideas in their ratings. Furthermore, the raters were permitted to take notes on the essay and/or the rubric pertaining to each essay. Additionally, they were told to feel comfortable to give feedback to the essays as if the papers would be returned to the students in order to encourage the raters to react to students' essays in a more authentic way. As for the time planning of individual essay scoring, the researcher did not limit the raters, considering that each rater should not feel pressure to allocate a necessary amount of time for scoring each essay.

Following the completion of scoring each essay, the raters provided written explanations for their ratings assigned to the essay. In other words, the raters were asked to justify their decisions about the essay by presenting simple reasons from their points of view about the students' written productions. In this way, the researcher aimed to obtain original ratings out of actual assessments and give an opportunity to the raters to consider their scores. These explanations were used to help find out any causal relationship between the assessment strategies employed, the scores assigned to the essays, and the aspects of the essays that the raters attended to in their evaluations. After finalizing the assessments, the raters returned the essay packs to the researcher.

Recording raters' spoken thoughts. As for the qualitative data, the participating raters were trained to utilize TAPs in which they were required to state out loud what they were thinking about the essays during their assessments. There were 16 pre-determined essays of different qualities in the data pack to be assessed using TAPs. The researcher included a reminder about TAPs on the top of the relevant essays in addition to specifying them in the

TAPs instructions list, aiming to prevent raters from thinking aloud about any other essays accidentally. Because the participant raters were full-time employees and they were not free-lance raters, the researcher thought that asking raters to assess all the essays that required thinking aloud at once and to record their voices in the same audio file would cause extreme data loss, in that it would inconvenience the raters. To this end, separate audio-recordings were demanded for each essay from the raters.

In order to examine the whole grading process that the raters went through, they were told to keep talking from the beginning of scoring to the completion of rating the essays. Being natural throughout this process was crucial for the reliability of data gathered from the recordings. Therefore, the researcher underscored the confidentiality of raters' identities in order to encourage them to speak naturally and continuously even if what they said might seem trivial. Additionally, the raters were reminded not to rationalize their ideas at length but to be natural as the purpose of the technique was to find out their natural thought processes when they were assessing the EFL essays. As regard to the language use during the TAP implementation, the researcher did not prioritize a specific language preference; instead, the raters were free to speak either English or Turkish or even both to elicit relevant data related to assessment strategies. Along with the essays and the rubrics that were used for each essay, the participants delivered the recordings to the researcher.

Data Preparation

The data for the thesis were collected through qualitative and quantitative methods. Each set of data were prepared for analysis using Microsoft Excel; however, different computer programs were used during the analysis. While SPSS Statistics 24.0 was used for descriptive and inferential statistics for qualitative and quantitative data sets, EduG software program was employed to carry out generalizability analysis based on quantitative data. Although several programs—GENOVA, ETUDGEN, SPSS, SAS, and MATLAB—can be used for G-theory analysis, the researcher preferred EDUG 6.0 for its user-friendly features (Güler et al., 2012).

Preparing quantitative data. The quantitative data set was comprised of scores that participant raters assigned to the essays. A total number of 9,900 essay scores (1,650 sub-scores for each of the components—grammar, content, organization, style and quality of expression, and mechanics as well as 1,650 total scores) obtained from 50 essays scored by 33 raters were recorded into Excel. Additionally, the scores assigned to the aforementioned essay components were summed up in the Excel program to double-check whether the total scores were calculated correctly by the raters.

Descriptive and inferential statistics. Descriptive and inferential statistics on SPSS were conducted in order to analyze whether there were any significant differences among raters in different experience groups in terms of the scores that they assigned to the low-quality and high-quality essays. These sets of analysis were carried out not only on total scores but also on the sub-scores assigned to different components of the essays. Additionally, descriptive statistics were conducted for the codes obtained from the analysis of TAPs and written explanations to compare the distribution of decision-making strategies across rater groups with the consideration of their experience level and the quality of the essays.

G-theory analysis. This study employed G-theory framework by using the computer program EduG in order to estimate the relative contributions of students, raters, and essay quality and their interactions to the variance in the essay scores. Additionally, generalizability and dependability coefficients were calculated in order to see whether the reliability of the scores assigned to low- and high-quality essays differ among rater groups with varying rating experience. In the present study, students were the object of measurement while essay quality and rater were considered random facets. Because all the students (persons as p) wrote the essays and all the participating raters (r) scored the essays in high and low qualities (q), the G-study design was completely crossed as *p x r x q*. Based on the aforementioned facets, a number of G-studies were conducted as follows:

a) Person-by-rater-by-quality (*p x r x q*) random effects G-study was conducted to obtain seven independent sources of variation including persons (*p*), rater (*r*), quality (*q*), person-by-rater (*p x r*), person-by-quality (*p x q*), rater-by-quality (*r x q*), and person-by-rater-by-quality (*p x r x q*) for 50 essays scored using analytic scoring methods.

   Additionally, generalizability and dependability coefficients were calculated for the reliability of the data set.

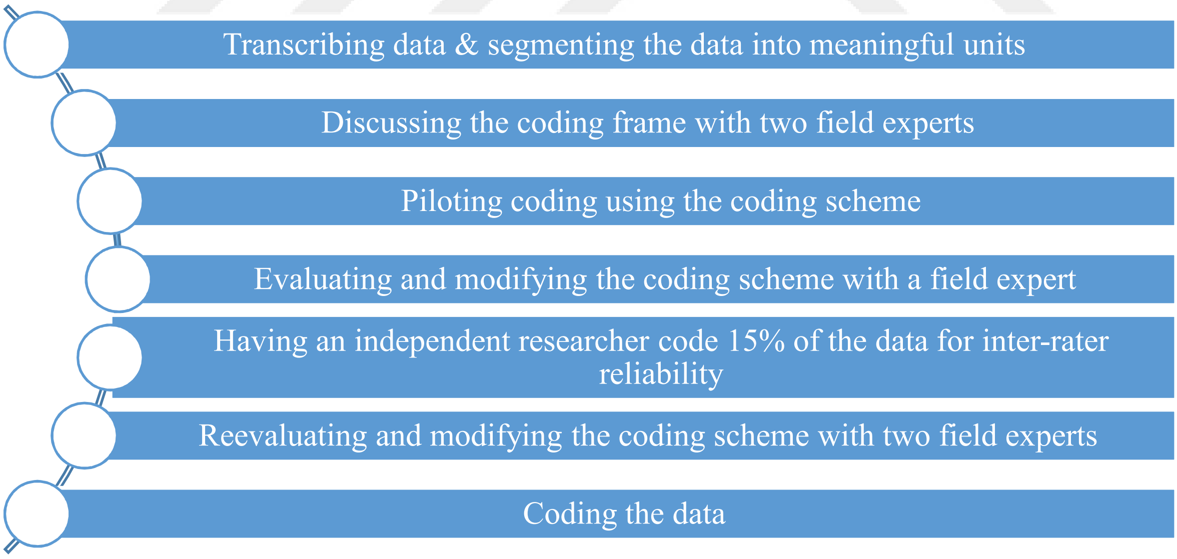
b) Person-by-rater (*p x r*) random effects G-study was conducted to obtain variance component estimates for three independent sources of variation including persons (*p*), rater (*r*), and person-by-rater (*p x r*) for 25 low-quality essays scored using analytic scoring methods. Additionally, generalizability and dependability coefficients were calculated for the reliability of the data set.

c) Person-by-rater (*p x r*) random effects G-study was conducted to obtain variance component estimates for three independent sources of variation including persons (*p*), rater (*r*), and person-by-rater (*p x r*) for 25 high-quality essays scored using analytic scoring methods. Additionally, generalizability and dependability coefficients were calculated for the reliability of the data set.

Moreover, the participant raters reported their previous rating experience and they were grouped into three categories based on the number of years they spent on assessing EFL writing previously. With this in mind, three experience groups were obtained: 1) raters with four years' or less experience, 2) raters with five to six years' experience, and 3) raters with seven years' or more experience. In order to compare generalizability and dependability coefficients of the ratings assigned by each rater experience group, person-by-rater-by-quality (*p x r x q*) random effects G-study for all essays and person-by-experience (*p x r*) random effects G-study for low- and high-quality essays were conducted.

Preparing qualitative data. The qualitative data set included TAPs that the raters provided for 16 of the essays during their assessments and the three reasons they reported for their assigned score to each of the 50 essays. While the data derived from TAPs are considered central to the qualitative aspect of the research, written explanations listed for the justification of assigned scores to the essays were used for triangulation purposes. The following sections provide detailed procedures for preparation and analysis of the two qualitative data sets.

Transcribing and coding think-aloud protocols. The data collected from the TAPs were analyzed using qualitative content analysis. In doing so, the researcher used a deductive approach, also known as top-down approach (Boyatzis, 1998), to analyze the data with the employment of a coding scheme (see Appendix I) adapted from Cumming et al. (2002). This process included the systematic planning of several phases as follows:



Transcribing data & segmenting the data into meaningful units

Discussing the coding frame with two field experts

Piloting coding using the coding scheme

Evaluating and modifying the coding scheme with a field expert

Having an independent researcher code 15% of the data for inter-rater reliability

Reevaluating and modifying the coding scheme with two field experts

Coding the data

Figure 8. Steps for qualitative content analysis.

The first step of the aforementioned process was to transcribe the TAPs that were collected from the participating raters. The total duration of the protocols were approximately

62 hours and 48 minutes. For high-quality essays, the raters recorded a total of 30 hours and 54-minute length verbal protocols with an average of 8:19 minutes per essay and 31 hours and 55-minute length protocols were recorded, resulting in an average of 8:35-minute length per low-quality essay. Following that, the researcher segmented the data into meaningful coding units, each of which can apply to the sub-categories of the coding frame. These meaningful units varied from a single word to a set of sentences that focused on the same aspect of the commented essay without any interruptions. Cumming et al. (2002, p. 76) relied on three criteria in order to segment the TAPs into meaningful and comparable units: "a) by pauses of 5 seconds or more, b) by the rater reading aloud a segment of the composition, or c) by the start or end of the assessment of a single composition". However, unlike in Cumming et al.'s study, the coding in the present study relied on a coding scheme in a deductive manner and the raters scored the essays one at a time instead of all at once, which means that they recorded a different audio file for each essay. In this regard, the researcher followed three criteria to divide the TAPs into meaningful units: a) by the rater reading the essay or a part of the essay, b) when the rater attend to the same aspect of the composition in their comments in a continuous manner, and c) by the rater stating a complete thought in a holistic manner.

After the data were segmented, each item of the coding scheme was discussed with two field experts considering a set of the transcribed qualitative data. Then, the researcher coded a number of fifty transcribed protocols using the coding frame followed by another expert consultation to evaluate and modify the sub-categories of the coding frame. It was found that while some coding categories did not appear in the trial-coding phase, new sub-categories were added to the frame and some of them were revised. For example, the category "scanning whole composition" was revised as "scanning or skimming whole composition" as a self-monitoring interpretation strategy, and "reading or interpreting the scoring scale" was found to be a new self-monitoring interpretation strategy. Barkaoui (2011b) identified the latter strategy in his study in which he employed the same coding frame developed by Cumming et al. (2002).

In addition, the researcher attended a conference (Şahan & Razı, 2017) to present the preliminary findings of the piloted qualitative data analysis that focused on the use of the coding frame on verbal protocols. In this conference, the researcher discussed the methodological aspects of this dissertation study with Alister Cumming, the lead author and one of the scholars who contributed to the development of the coding scheme used in this study. A discussion was maintained on revised and new strategies obtained in the pilot analysis. It was agreed that cultural and contextual differences would be an important consideration in adapting the sub-categories of the coding frame.

In order to check the inter-rater reliability of the coding, one independent researcher, who had expertise in transcribing and analyzing qualitative data including deductive and inductive coding, coded a random sample of 15% of the TAPs (Barkaoui, 2007b; Lumley, 2002). The statistical analysis on the similarity of coding carried out by the researchers revealed a very good agreement between the two coders, $\kappa = .83$ with $p < .001$. A value of over .80 represents very good agreement (Landis & Koch, 1977).

After inter-rater reliability was ensured between the researchers, a follow-up discussion was carried out to elaborate into the reasons for disagreement where the researchers differed in their coding. To do so, each sample of transcriptions coded by the two researchers independently was examined to determine which strategies in the coding scheme resulted in disagreement between the raters. Following the piloting and inter-rater reliability phases, the researcher made the following changes by consulting two independent experts:

- Firstly, one of the interpretation strategies, “read or reread composition” under the self-monitoring focus was revised as “read or reread text” since it was found that raters sometimes (re)read one part of the essay when they had difficulty in interpreting the text.

- Secondly, one of the interpretation strategies, "scan whole composition" under the self-monitoring focus was changed to scan or skim composition after the pilot analysis. Following the coding process for inter-rater reliability, this strategy was redefined as "scan or skim text" because the raters in this study scanned or skimmed a part of the composition at times.
- Thirdly, following the pilot analysis of TAPs, a new self-monitoring interpretation strategy—"read or interpret scoring scale"—was added to the scheme, which was also identified by Barkaoui (2007b).
- Fourthly, one of the judgement strategies, "consider own personal response or biases" was revised and changed to "consider own personal response, expectations or biases" given that the two coders had problems in finding a matching strategy when raters talked about their personal expectations from the essay or a segment of the essay.
- The fifth revision made on the coding scheme was to merge the two judgement strategies under rhetorical and ideational focus: "assess task completion" and "assess relevance". This created a single category called, "assess tasks completion and relevance". Because in the assessment context of this study, uncompleted essays were detected and discarded during essay quality divisions, it was decided that raters' comments about task completion during their think-aloud assessments refereed to the students' responses to the topic in terms of relevancy, thus justifying the merger of the two categories.
- Another change was made to one of the judgements strategies within the language focus, which deals with the quantity of the written production. The

strategy was originally, "assess quantity of total written production", and it was modified as, "assess quantity of written production", as raters sometimes considered the quantity of the text at the sentence or paragraph level as well in the TAPs.

- A final point was made on the punctuation used while transcribing the voice recordings in that it was agreed that in one sentence, there might be more than one meaningful unit that might indicate different strategies in the coding scheme. In this sense, coding and segmentation were not bound to the punctuation of the transcription.

Following the revisions made to the coding scheme, a final discussion was carried out about coding the TAPs and the two researchers, who had almost perfect agreement ($\kappa = .83$), carried on coding a number of 446 TAPs. In doing so, they completed coding the TAPs individually, and when any kind of uncertainty occurred, the researcher made the final decision following negotiations with the other coder.

Thematic content analysis for written score explanations. The researcher went through all the essays ($N = 1650$) to analyze the written score explanations that raters provided to each essay. In doing so, the explanations, which generally appeared as short statements including an aspect of the essay described by an adjective (e.g. good grammar, fair language use, satisfactory content, etc.) were analyzed using thematic content analysis. Each explanation for the assigned score was coded in terms of *focus* (e.g. grammar, content, topic development) as well as *type* (e.g. positive or negative) adapting the coding system developed by Barkaoui (2010c). It should be noted that neutral explanations did not occur in the data since the researcher had told the raters to be clear in their explanations regarding the type. The purpose of this analysis was to determine the frequency of each theme that raters attended to during their assessments to triangulate the data derived from the TAPs. Inter-rater reliability was ensured with the help of an independent researcher, who is a doctoral student in the field of language education. The independent researcher coded 10% of the data using the same inductive techniques of theme identification and connotation categorization. The statistical analysis on the similarity of coding carried by the researchers revealed a very good agreement between the two coders, $\kappa = .89$ with $p < .001$.

Research Ethics

There are no foreseeable risks associated with participation in this research project, and the researcher fulfilled the requirements of research ethics in every phase of the data collection and thesis writing process. First, the students (18+) whose essays were used in the study were informed about the purpose of the research and their consent was received before the compositions were collected. Second, official permission was received from the Dean's Office to which the students were enrolled to collect the student essays for research purposes (Appendix K). Third, three expert raters were contacted to make the quality division of the essays and each consented to participate in the research. Fourth, the raters were contacted through email and asked for their voluntarily participation in the research. Overall, all participants were ensured that the participation is voluntary, their identities are confidential, and they may withdraw from the study without any penalty. Furthermore, the researcher received official permissions from the raters' institutions for their participation in the research (Appendix L). Finally, the three expert raters and 33 participant raters were informed that they would be compensated for their efforts if the researcher received funding from domestic or international research foundations.

Summary

This chapter explained the methodological aspects of the current research study. First, G-theory as a theoretical framework was explained followed by a detailed description of the demographics of the participants. In the following sections, data collection instruments were presented and the adaptation processes of analytic rating scale and data coding scheme with the pilot analysis outcomes were provided. Next, quantitative and qualitative data collection procedures were detailed respectively. In addition, the chapter continued with the steps in which data were prepared for analysis and the computer programs used for analyzing the data were introduced in accordance with the type of the analysis. Finally, ethical issues in educational research were explained. The findings and results of the study are presented in the next chapter.

Chapter IV

Results

This chapter presents the results of each research question separately along with their respective data analysis results. As a convergent parallel design as a mixed-methods approach was used in the study, the results pertaining to quantitative and qualitative data are analyzed and presented separately. The chapter starts with the presentation of the quantitative data analysis results in which four research questions are answered. Following that, the qualitative data analysis results are provided to answer the remaining two research questions.

Quantitative Data Analysis Results

As for the analysis of quantitative data, SPSS 24.0 was used for descriptive and inferential statistics to address the first two research questions (RQ). **RQ1** inquired *whether there were any significant differences between the analytic scores assigned to high-quality and low-quality essays* while **RQ2** asked *whether there were any significant differences among the analytic scores assigned by raters with varying experience*. While answering these questions, descriptive and inferential statistics were conducted with specific emphasis on essay quality and previous rating experience. G-theory analysis was conducted to answer **RQ3,** which aimed to *explore the sources of score variation that contribute most (relatively) to the score variability of the analytic scores of EFL essays,* and **RQ4,** which asked *whether the reliability of the analytic scores of raters differ based on their amount of experience*.

Results for RQ1. The first research question is: Are there any significant differences among the analytic scores assigned to the low- and high-quality EFL essays? In order to answer this research question, descriptive and inferential statistics were conducted and the results are presented using both figures and tables.

Figure 9 and 10 display the deviation of median values as well as ranges for the essay scores assigned to high- and low-quality essays. Boxplots are considered a good way to visualize the distribution of data because they divide the data into quartiles. The body of the boxplot, which is shown as the colored box in Figures 9 and 10, represents 50% of the data. The horizontal black line in the box shows the median of the data set; this also represents the second quartile (Q2). The vertical lines extending from the top to the bottom of the box are known as whiskers and represent the remaining 50% of the data. The line extending below the box includes the data from the smallest non-outlier to the first quartile (Q1) of the data set. The whisker extending above the box includes data points from the third quartile (Q3) to the largest non-outlier. Because the boxplot includes data within a 95% confidence interval, outliers are plotted as points on the graph above or below the boxplot. The vertical distance of the boxplot represents the range, and the length of the quartiles in the boxplot illustrates the skewness patterns of the data. While the first boxplot graph depicts the distribution of the scores assigned to the high-quality papers (*n* = 25), the second boxplot graph shows the distribution of the scores assigned to the low-quality papers (*n* = 25).

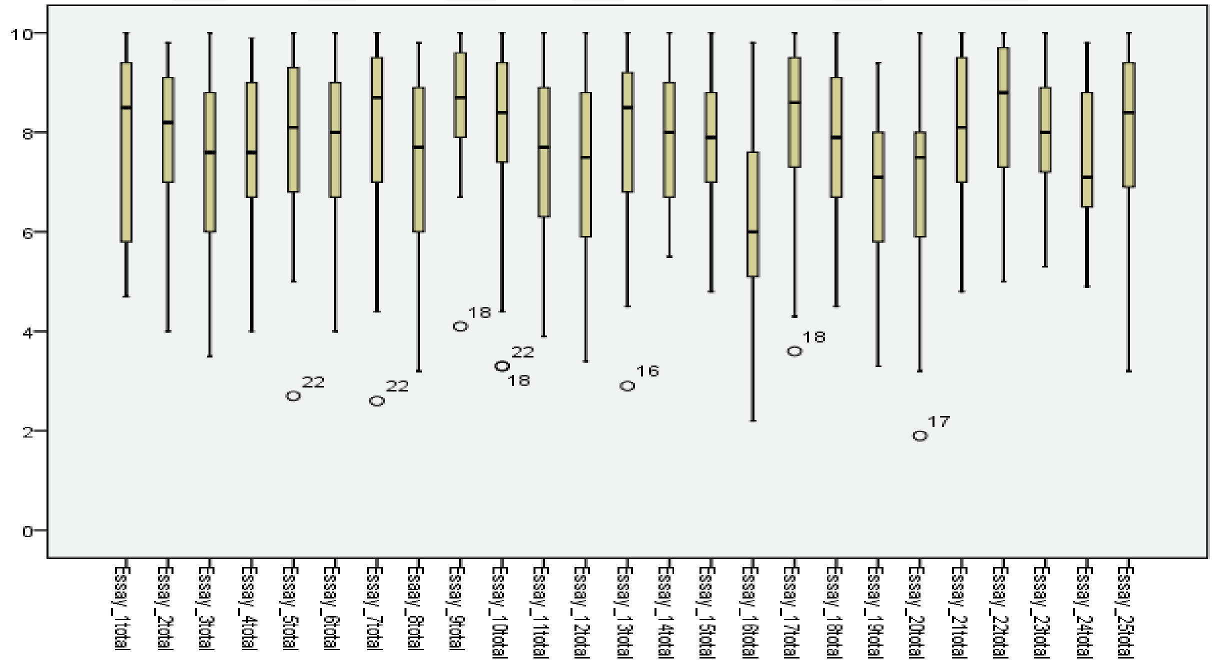


Figure 9. Boxplots for the total scores assigned to high-quality essays.

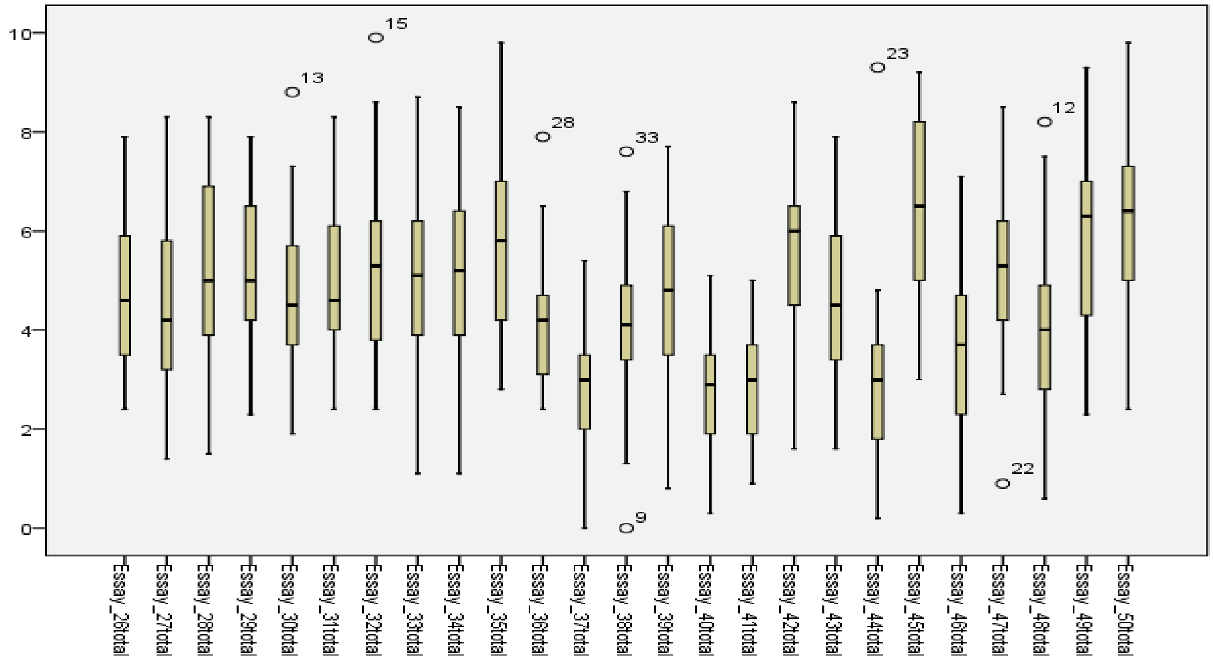


Figure 10. Boxplots for the total scores assigned to low-quality essays.

As can be seen in the boxplots graphs, for high-quality essays, smaller ranges of essay scores can be observed for the first two quartiles and larger ranges can be observed for the second two quartiles, suggesting that scores were concentrated on the higher end of the rubric. For low-quality essays, in contrast, the range of scores appears to be more evenly distributed across the four quartiles. Additionally, there appears to be more variance in the median scores for low-quality papers as compared to high-quality papers, for which median scores appeared relatively similar across the 25 essays.

The data were further investigated through descriptive and inferential statistics in order to have a better idea about the distribution of the scores and the differences between two sets of essays. Because the boxplots suggested great ranges for both high-quality and low-quality essays, the range between minimum and maximum scores was calculated for each essay (see Appendix M for high-quality essays and Appendix N for low-quality essays). Overall, the mean range for all essays was 6.39 and the mean range for high-quality essays was 6.13 while the mean range for low-quality essays was 6.66. These values indicated a striking range in the scores out of a 10-point scale.

In order to compare the differences between the scores assigned to high-quality and low-quality essays, an independent-samples *t*-test was conducted. The analysis revealed a significant difference in the scores assigned by raters to high- and low-quality essays. The scores assigned by raters differed significantly between the high-quality (*M* = 7.67, *SD* = 0.49) and low-quality essays (*M* = 4.65, *SD* = 1.09; *t*(33.5) = 12.55, *p* < .001, two-tailed). The magnitude of the differences in the means (mean difference = 3.02, 95% *CI*: 2.53 to 3.51) was very large (eta squared = 0.87). Furthermore, independent-samples *t*-tests were conducted to compare the scores assigned to high- and low-quality essays by each rater. For each rater, the results yielded significant differences in the scores assigned to high- and low-quality essays. For 32 out of 33 raters the significance value was found to be *p* < .001; for one rater the significance was calculated as *p* = .005. These results suggest that there are significant differences in the analytic scores assigned to low- and high-quality EFL essays.

Results for RQ2. The second research question is: Are there any significant differences among the analytic scores assigned by raters with varying previous rating experience?

As mentioned in the demographics section, the raters were divided into three categories based on their previous rating experience. Raters who reported four years or less scoring experience were categorized in the low-experienced group (*n* = 13); raters with five to six years of rating experience were labelled as medium-experienced raters (*n* = 10) while raters with seven years' or more experience fell into the high-experienced rater category (*n* = 10). In order to answer RQ2, the scores assigned to high- and low-quality essays were compared using descriptive and inferential statistics. Figure 11 shows the mean essay scores for each of the 25 high-quality essays according to rater experience.

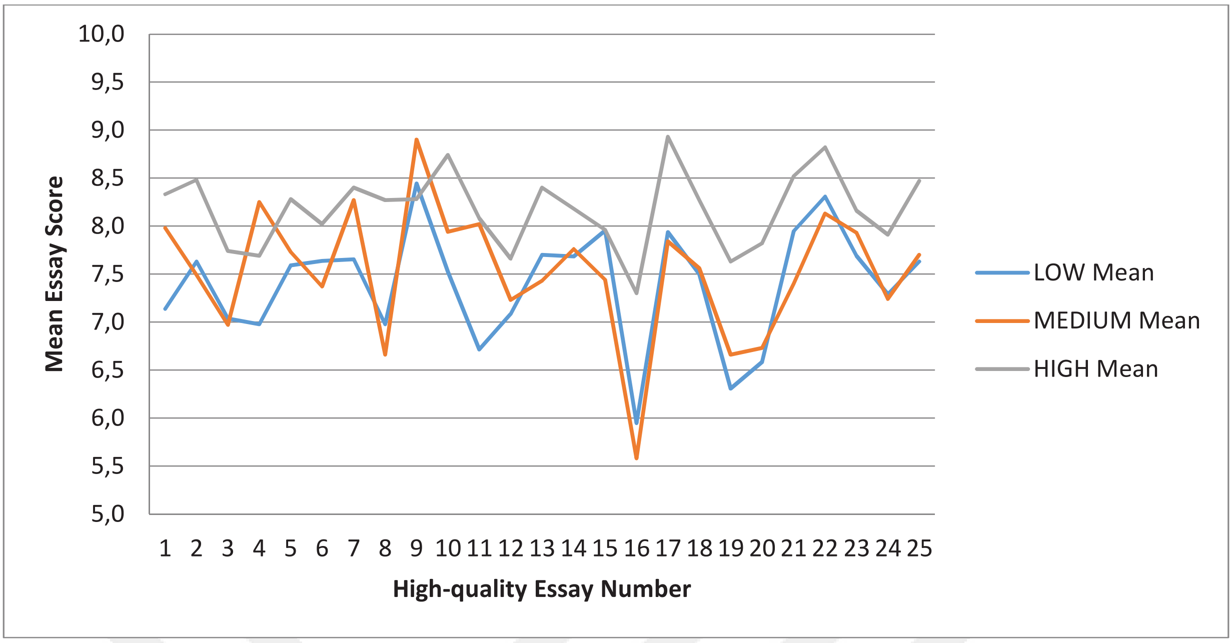


Figure 11. Scoring trend for high-quality essays based on rater experience.

As can be seen from Figure 11, raters in the more experienced group tended to give higher scores to high-quality essays than raters with less experience, including raters in both the low- and medium-experienced groups. Figure 12 shows the mean essay score for each of the 25 low-quality essays according to rater experience.

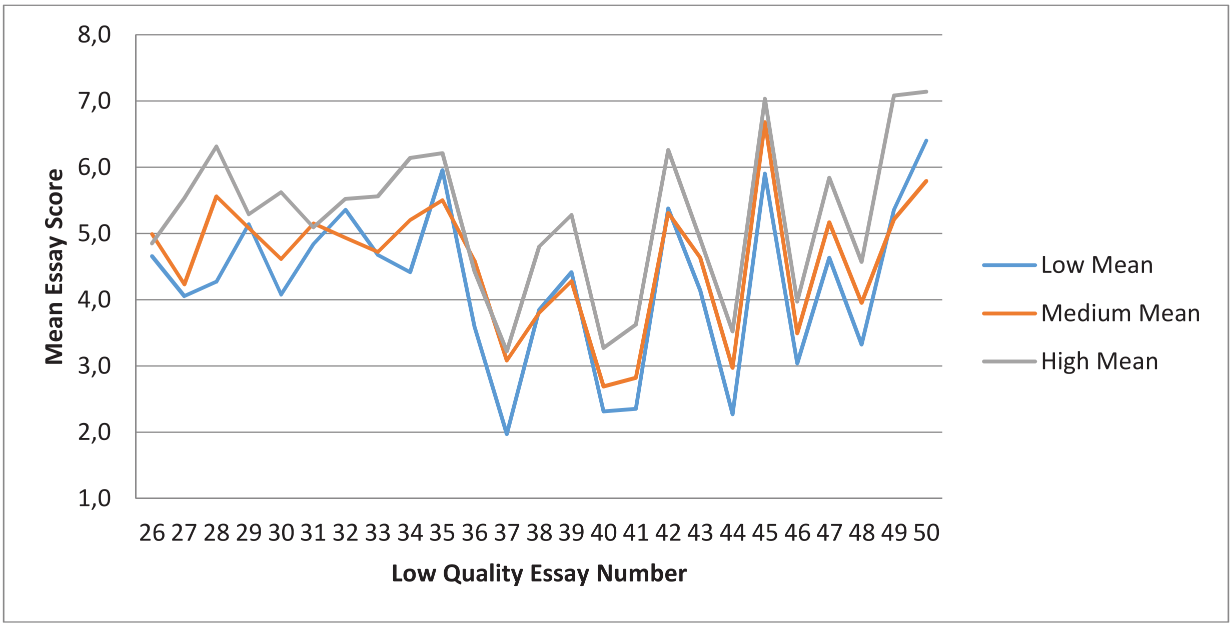


Figure 12. Scoring trend for low-quality essays based on rater experience.

As evident from Figure 12, the same trend that was observed for high-quality essays can also be seen for low-quality essays: raters with more experience tended to give higher scores than raters with less experience. Moreover, Figure 12 shows that a wide range of mean scores was given to the low-quality essays, with the highest mean score of 7.14 given to essay number 50 by the high experience group and the lowest mean score of 1.97 given to essay 37 by the low experience group. Table 10 shows the mean essay scores given for high- and low-quality essays according to experience group.

Table 10

Mean Essay Scores by Experience Groups

		Mean score	
		High-quality essays	Low-quality essays
Experience	Low	7.39	4.25
	Medium	7.52	4.57
	High	8.17	5.24

As can be seen in Table 10, the high-experienced group tended to give higher scores to both high- and low-quality essays, while the low-experienced group tended to give lower scores to both high- and low-quality essays. The mean scores assigned to individual essays by each rater group can be found in the Appendix pages (Appendix O for high-quality essays and Appendix P for low-quality essays).

After observing general trends in the data, statistical analyses were carried out using SPSS 24 to examine whether the trends observed in the data were statistically significant. Namely, analyses were conducted to examine whether the tendency of more experienced raters to assign higher scores to both high- and low-quality essays was statistically significant. To do this, non-parametric tests were conducted to compare the means across three groups (Gr1, $n_{\text{low-experienced}} = 13$; Gr2, $n_{\text{medium-experienced}} = 10$; Gr3, $n_{\text{high-experienced}} = 10$).

A Kruskal-Wallis test revealed no statistically significant differences in the mean scores assigned to high-quality essays across three experience groups (Gr1, $n_{\text{low-experienced}} = 13$; Gr2, $n_{\text{medium-experienced}} = 10$; Gr3, $n_{\text{high-experienced}} = 10$), χ^2 (2, $n = 33$) = 2.74, $p > .05$. The high-experienced group recorded a higher median score (*Mdn* = 8.60) than the medium-experienced and low-experienced groups, which recorded median values of 7.79 and 7.36, respectively.

A Kruskal-Wallis test revealed a statistically significant difference in the mean scores assigned to low-quality essays across three experience groups (Gr1, $n_{\text{low-experienced}} = 13$; Gr2, $n_{\text{medium-experienced}} = 10$; Gr3, $n_{\text{high-experienced}} = 10$), χ^2 (2, $n = 33$) = 6.72, $p = .04$. The high-experienced group recorded a higher median score (*Mdn* = 5.28) than the other two groups, which recorded median values of 4.28 for the low-experienced group and 4.38 for the medium-experienced group.

Following the findings of the Kruskal-Wallis test, follow-up Mann-Whitney *U* tests were performed to determine which of the groups were statistically significant from each other. A Mann-Whitney *U* test revealed statistically significant differences between low- (*Mdn* = 5.28, $n = 13$) and high- (*Mdn* = 4.28, $n = 10$) experienced groups, $U = 23$, $z = -2.61$, $p = .01$, $r = .54$. Mann-Whitney *U* tests revealed no statistically significant differences for the other paired groups.

Following the Kruskall-Wallis test which revealed significant differences between the mean scores assigned to low-quality essays across groups, Kruskall-Wallis tests were carried out on each low-quality essay ($n = 25$) to reveal significant differences within the scores assigned to each individual essay across experience groups. The Kruskall-Wallis tests revealed statistically significant differences in the scores assigned to the following three essays: Essay 28 (χ^2 (2, $n = 33$) = 7.36, $p = .03$), Essay 41 (χ^2 (2, $n = 33$) = 6.27, $p = .04$), and Essay 49 (χ^2 (2, $n = 33$) = 7.19, $p = .03$). For each essay determined to have received statistically significant scores across groups, follow up Mann-Whitney *U* tests were conducted.

For Essay 28, a Mann-Whitney *U* test revealed statistically significant differences between the low- (*Mdn* = 3.90, *n* = 13) and high- (*Mdn* = 6.60, *n* = 10) experienced groups, *U* = 27, *z* = -2.36, *p* = .02, *r* = .49. A Mann-Whitney *U* test also revealed statistically significant differences between the low- (*Mdn* = 3.90, *n* = 13) and medium- (*Mdn* = 5.25, *n* = 10) experienced groups, *U* = 31.5, *z* = -2.08, *p* = .04, *r* = .43.

For Essay 41, a Mann-Whitney *U* test revealed statistically significant differences between the low- (*Mdn* = 2, *n* = 13) and high- (*Mdn* = 3.70, *n* = 10) experienced groups, *U* = 28, *z* = -2.30, *p* = .02, *r* = .48.

For Essay 49, a Mann-Whitney *U* test statistically significant differences between the low- (*Mdn* = 5, *n* = 13) and high- (*Mdn* = 7, *n* = 10) experienced groups, *U* = 28, *z* = -2.30, *p* = .02, *r* = .48. A Mann-Whitney *U* test also revealed statistically significant differences between the medium- (*Mdn* = 4.45, *n* = 10) and high- (*Mdn* = 7, *n* = 10) experienced groups, *U* = 19, *z* = -2.35, *p* = .02, *r* = .53. Mann-Whitney *U* tests revealed no statistically significant differences for the other paired groups for each of these three essays.

In addition to the analyses investigating variations in the total scores assigned to essays, the scores assigned to each component of the rubric (grammar, content, organization, style and quality of expression, and mechanics) were analyzed to examine whether statistically significant differences could be found in the scores assigned by raters in low-, medium- and high-experienced groups to each component. First, non-parametric tests were conducted to compare the means assigned to component scores across three rater experience groups (Gr1, $n_{\text{low- experienced}}$ = 13; Gr2, $n_{\text{medium-experienced}}$ = 10; Gr3, $n_{\text{high-experienced}}$ = 10).

A Kruskal-Wallis test revealed statistically significant differences in the mean scores assigned to the mechanics component of low-quality essays across three experience groups (Gr1, $n_{\text{low-experienced}}$ = 13; Gr2, $n_{\text{medium-experienced}}$ = 10; Gr3, $n_{\text{high-experienced}}$ = 10), χ^2 (2, *n* = 33) = 6.06, *p* = .048. For the mechanics component of the low-quality essays, the high-experienced group recorded a higher median score (*Mdn* = 0.68) than the low-experienced group (*Mdn* = 0.50) and the medium-experienced group (*Mdn* = 0.58). No significant differences between groups were found for the other components of low-quality essays. Overall, no significant differences were found between groups for the scores assigned to any of the components for high-quality essays.

Following the findings of the Kruskal-Wallis test, follow-up Mann-Whitney *U* tests were performed to determine which of the experience groups were statistically significant from each other. A Mann-Whitney *U* test revealed statistically significant differences between the low- (*Mdn* = 0.50, *n* = 13) and high- (*Mdn* = 0.58, *n* = 10) experienced groups, *U* = 29.5, *z* = -2.20, *p* = .03, *r* = .46). Mann-Whitney *U* tests revealed no statistically significant differences for the other paired groups.

In conclusion, statistically significant differences were found between the scores assigned to low-quality essays by high-experienced and low-experienced rater groups. Further analysis revealed that these two experience groups significantly differed in their mechanics component scores. No statistically significant differences were found between rater groups for their total scores or component scores given to high-quality essays.

Comparison of the scores based on self-described experience. The findings derived from the differences between the scores assigned to the essays by the raters pertaining to each experience group encouraged the researcher to conduct further analyses based on the experience levels that raters self-described. In order to investigate the extent to which the distribution of raters into groups according their self-described experience corresponded to the categorization of raters according to their reported experience in rating papers, descriptive statistics were conducted to the determine the overlaps between actual experience and self-described experience. Table 11 compares raters according to their self-described experience and their reported experience.

Table 11

Demographics of Participants Based on Their Reported and Self-described Experience in EFL Writing Assessment

		Self-described Experience					
		No experience	*Little experience*	*Some experience*	*Experienced*	*Very experienced*	***Total***
Exp. Group	*Low*	2	3	4	4	0	**13**
	Medium	0	1	5	4	0	**10**
	High	1	0	5	3	1	**10**
Total		**3**	**4**	**14**	**11**	**1**	**33**

As can be seen from Table 11, there was variation in the grouping of raters when raters were categorized according to their self-described experience rather than reported experience. Of the 13 raters in the low-experienced group, four described themselves as high-experienced raters and four described themselves as medium- (somewhat) experienced raters; only five of the 13 low-experienced raters self-identified as low-experienced raters. Additionally, five of the 10 high-experienced raters described themselves as medium- (somewhat) experienced raters, and one of the high-experienced raters described him or herself as a low-experienced rater. Among the raters who self-described as high-experienced (*n* = 12), there was an equal distribution among experience groups: four raters belonged to the low-experienced group, four belonged to the medium-experienced group, and four belonged to the high-experienced group. A similar pattern can be observed in the self-described medium-experienced group, of which four raters belonged to the low-experienced group, five belonged to the medium-experienced group, and five belonged to the high-experienced group. Thus, the grouping of raters according to self-described experience differed from the grouping according to reported experience in terms of years rating EFL essays. Raters did not necessarily self-identify with their groupings in terms of experience in years. Given the difference between these two sets of groupings, data were analyzed according to self-described experience in order to investigate whether a rater's perceived or self-described experience affect their rating behavior.

Firstly, trends in the mean essay scores assigned to high- and low-quality essays were examined with respect to raters' self-described level of experience rating papers. Raters were divided into three groups (low, medium, high) based on their self-described experience. Figure 13 shows the mean essay scores assigned to the 25 high-quality essays according to raters' self-described experience.

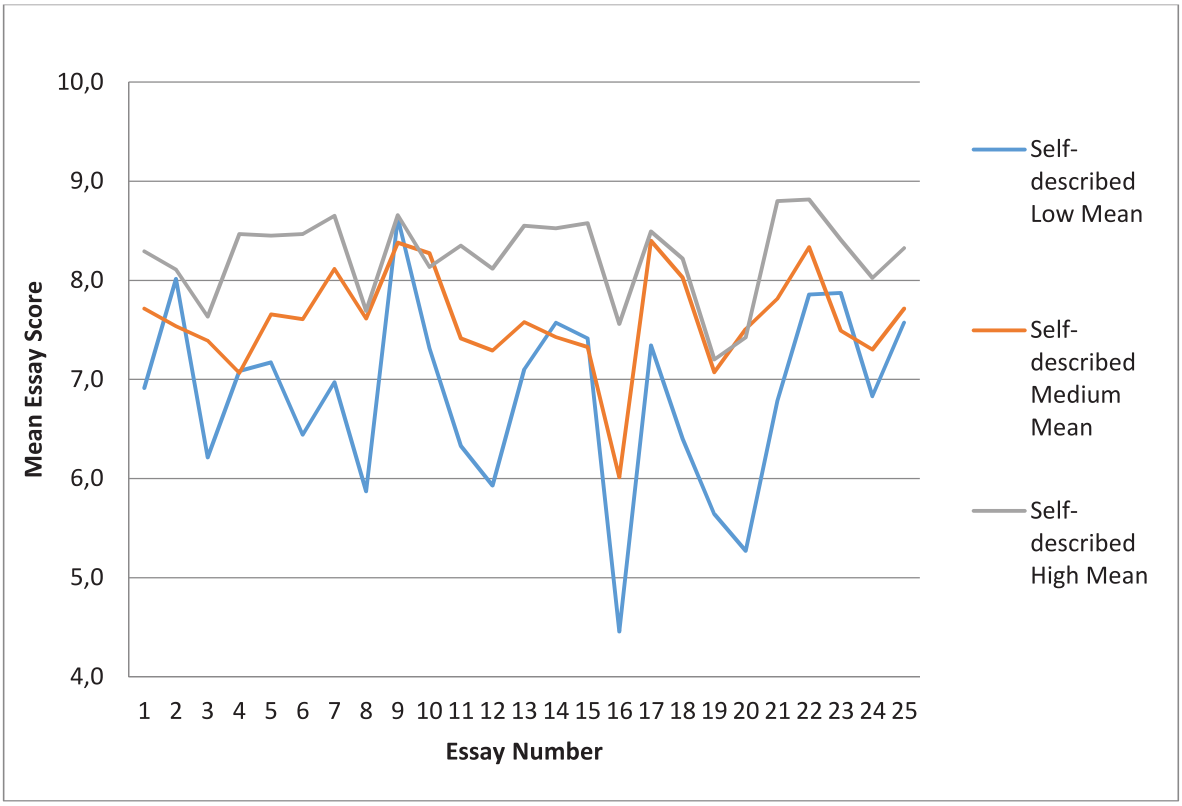


Figure 13. Scoring trend for high-quality essays based on self-described experience.

Figure 13 demonstrates that a similar trend can be seen in the mean scores assigned to high-quality essays by self-described rater experience groups as was observed with reported experience groups. Raters who describe themselves as having more experience tended to assign higher scores to high-quality essays than raters who described themselves as having less experience (self-described low- or self-described medium-experienced groups). The data were similarly examined for low-quality essays. Figure 14 shows the mean scores assigned to each of the 25 low-quality essays according to raters' self-described experience level.

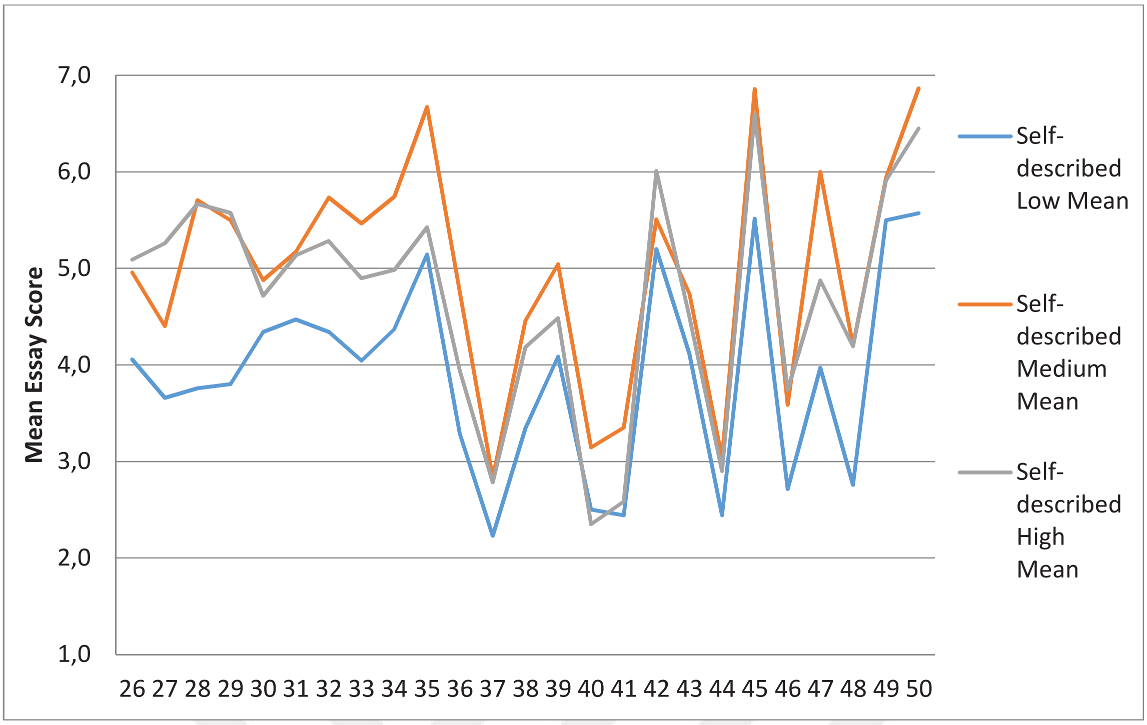


Figure 14. Scoring trend for low-quality essays based on self-described experience.

Figure 14 suggests that raters' with less self-described experience tended to assign lower scores than raters with more experience (self-described high- or self-described medium-experienced groups). However, for low-quality essays, the trend among self-described high-experienced raters to assign higher scores is less clear. Rather, self-described high- and self-described medium- experienced raters tended to assign similar scores, with self-described medium-experienced raters sometimes assigning higher scores than their self-described high-experienced peers. Table 12 summarizes the mean scores given to high- and low-quality essays by raters according to their self-described experience.

Table 12

Mean Essay Scores by Self-described Experience Group

		Mean Score	
		High-quality essays	Low-quality essays
Self-described Experience	Low	6.84	3.91
	Medium	7.60	4.98
	High	8.24	4.70

As can be seen in Table 12, the high-experienced group tended to give higher scores to high- and low-quality essays compared to the low-experienced group, while the low-experienced group tended to give lower scores to both high- and low-quality essays. Medium-experienced raters, however, tended to give higher scores to low-quality essays than the high-experienced group of raters did.

After observing general trends in the data, statistical analyses were carried out using SPSS 24 to examine whether the trends observed in the data were statistically significant. Namely, analyses were conducted to examine whether the tendency of self-described high-experienced raters to assign higher scores to both high- and low-quality essays was statistically significant. To do this, non-parametric analyses were conducted to determine whether statistically significant differences could be found between the essay scores assigned by raters according to their self-described experience levels, as opposed to their reported experience in years rating EFL essays. Using Kruskal-Wallis tests, the mean scores assigned to high- and low-quality essays were compared across three self-described groups (Gr1, $n_{\text{self-described low-experienced}} = 7$; Gr2, $n_{\text{self-described medium-experienced}} = 14$; Gr3, $n_{\text{self-described high-experienced}} = 12$).

A Kruskal-Wallis test revealed statistically significant differences in the mean scores assigned to high-quality essays across three self-described experience groups (Gr1, $n_{\text{self-described low-experienced}} = 7$; Gr2, $n_{\text{self-described medium-experienced}} = 14$; Gr3, $n_{\text{self-described high-experienced}} = 12$), χ^2 (2, n = 33) = 7.23, p = .03. The self-described high-experienced group recorded a higher median score (*Mdn* = 8.33) than the self-described low-experienced group (*Mdn* = 6.87) and the self-described medium-experienced group (*Mdn* = 7.36).

Mann-Whitney *U* tests were then performed to determine which of the groups were statistically significant from each other. A Mann-Whitney *U* test revealed statistically significant differences between the self-described low- (*Mdn* = 6.87, n = 7) and the self-described high- (*Mdn* = 8.33, n = 12) experienced groups, U = 13, z = -2.45, p = .01, r = .56.

Mann-Whitney *U* tests revealed no statistically significant differences for the other paired groups.

No statistically significant differences were found in the mean scores assigned to low-quality essays across the self-described experience groups, χ^2 (2, *n* = 33) = 5.18, *p* = .08, although the self-described high-experienced group recorded a higher median score (*Mdn* = 5.04) than the two other groups (self-described low, *Mdn* = 4.26*;* self-described medium, *Mdn* = 4.69).

Following the Kruskall-Wallis tests on the mean scores given to high- and low-quality essays according to self-described experience groups, additional Kruskall-Wallis tests were carried out on each of the high-quality essays (*n* = 25) to reveal significant differences in the scores assigned to each individual essay across self-described experience groups. The Kruskall-Wallis tests revealed statistically significant differences in the scores assigned to eight essays, which are presented in Table 13.

Table 13

Kruskall-Wallis Test Results for High-quality Essays

Essay	χ^2 **(2, *n* = 33)**	***p***	***Self-described low (Mdn)***	***Self-described medium (Mdn)***	***Self-described high (Mdn)***
Essay 4	6.31	.04	7.20	6.85	9.05
Essay 6	6.84	.03	6.70	7.55	8.95
Essay 11	6.70	.04	6.50	7.45	8.60
Essay 12	6.37	.04	5.10	7.55	8.15
Essay 15	6.69	.04	7.30	7.85	8.65
Essay 16	11.03	.004	5.10	5.80	8.05
Essay 18	6.66	.04	6.30	7.95	8.20
Essay 21	7.58	.02	7.00	7.90	8.95

For each of the eight essays determined to have received statistically significant scores across self-described experience groups, follow up Mann-Whitney U tests were conducted.

For Essay 4, a Mann-Whitney U test revealed statistically significant differences between the self-described medium- ($Mdn = 6.85$, $n = 14$) and the self-described high- ($Mdn = 9.05$, $n = 12$) experienced groups, $U = 36$, $z = -2.47$, $p = .01$, $r = .49$.

For Essay 6, a Mann-Whitney U test revealed statistically significant differences between the self-described low- ($Mdn = 6.70$, $n = 7$) and the self-described high- ($Mdn = 8.95$, $n = 12$) experienced groups, $U = 13$, $z = -2.46$, $p = .01$, $r = .56$.

For Essay 11, a Mann-Whitney U test revealed statistically significant differences between the self-described low- ($Mdn = 6.50$, $n = 7$) and the self-described high- ($Mdn = 8.60$, $n = 12$) experienced groups, $U = 13$, $z = -2.45$, $p = .01$, $r = .56$.

For Essay 12, a Mann-Whitney U test revealed statistically significant differences between the self-described low- ($Mdn = 5.10$, $n = 7$) and the self-described high- ($Mdn = 8.15$, $n = 12$) experienced groups, $U = 17$, $z = -2.12$, $p = .03$, $r = .49$.

For Essay 15, a Mann-Whitney U test revealed statistically significant differences between the self-described medium- ($Mdn = 7.85$, $n = 14$) and the self-described high- ($Mdn = 8.65$, $n = 12$) experienced groups, $U = 37$, $z = -2.42$, $p = .02$, $r = .48$.

For Essay 16, a Mann-Whitney U test revealed statistically significant differences between the self-described low- ($Mdn = 5.10$, $n = 7$) and the self-described high- ($Mdn = 8.05$, $n = 12$) experienced groups, $U = 7.5$, $z = -2.92$, $p = .004$, $r = .67$. A Mann-Whitney U test also revealed statistically significant differences between the self-described medium- ($Mdn = 5.80$, $n = 14$) and the self-described high- ($Mdn = 8.05$, $n = 12$) experienced groups, $U = 40.5$, $z = -2.24$, $p = .03$, $r = .44$.

For Essay 18, a Mann-Whitney U test revealed statistically significant differences between the self-described low- ($Mdn = 6.30$, $n = 7$) and the self-described high- ($Mdn = 8.20$, *n* = 12) experienced groups, *U* = 15, *z* = -2.28, *p* = .02, *r* = .52. A Mann-Whitney *U* test also revealed statistically significant differences between the self-described low- (*Mdn* = 6.30, *n* = 7) and the self-described medium- (*Mdn* = 7.95, *n* = 14) experienced groups, *U* =18.5, *z* = -2.28, *p* = .02, *r* = .50.

For Essay 21, a Mann-Whitney *U* test revealed statistically significant differences between the self-described low- (*Mdn* = 7, *n* = 7) and the self-described high- (*Mdn* = 8.95, *n* = 12) experienced groups, *U* = 11.5, *z* = -2.59, *p* = .01, *r* = .59. Mann-Whitney *U* tests revealed no statistically significant differences for the other paired groups for each of the eight high-quality essays.

In order to further investigate the aforementioned findings, non-parametric tests were also conducted to compare the mean scores assigned to rubric components across three self-described rater experience groups (Gr1, $n_{\text{self-described low-experienced}}$ = 7; Gr2, $n_{\text{self-described medium-experienced}}$ = 14; Gr3, $n_{\text{self-described high-experienced}}$ = 12). Analyses were conducted on the scores assigned to five rubric components: grammar, content, organization, style and quality of expression, and mechanics.

A Kruskal-Wallis tests revealed statistically significant differences in the mean scores assigned to each of the five rubric components for high-quality essays across the three self-described groups (Gr1, $n_{\text{self-described low-experienced}}$ = 7; Gr2, $n_{\text{self-described medium-experienced}}$ = 14; Gr3, $n_{\text{self-described high-experienced}}$ = 12:). Table 14 summarizes the findings for each rubric component of the high-quality essays:

Table 14

Kruskall-Wallis Test Results for High-quality Essay Component Scores by Self-described Experience Groups

Component	**χ^2 (2, n = 33)**	***p***	***Self-described low (Mdn)***	***Self-described medium (Mdn)***	***Self-described high (Mdn)***
Grammar	8.89	.01	1.10	1.24	1.37
Content	6.38	.04	1.83	2.13	2.41
Organization	7.44	.02	1.64	1.83	2.09
Style & Quality of Expression	6.47	.04	1.40	1.52	1.65
Mechanics	6.35	.04	0.79	0.91	0.92

For every component, the self-described high-experienced group recorded higher median scores than the self-described low- and the self-described medium-experienced groups for each component (grammar, content, organization, style & quality of expression, and mechanics). The self-described low-experienced group recorded lower median scores than both the self-described medium- and the self-described high-experienced groups for each component.

Mann-Whitney *U* tests were performed to determine which of the self-described experience groups were statistically significant from each other. Table 15 summarizes the statistically significant differences between groups for the rubric components for high-quality essays.

Table 15

Mann-Whitney U Test Results for High-quality Essay Component Scores by Self-described Experience Groups

Component	**Groups**	***n***	***Mdn***	***U***	***z***	***p***	***r***
Grammar	Low	7	1.10	11	-2.62	.009	.60
	High	12	1.37				
Grammar	Medium	14	1.24	45	-2.01	.045	.39
	High	12	1.37				
Content	Low	7	1.83	16	-2.20	.03	.50
	High	12	2.41				
Organization	Low	7	1.64	14	-2.37	.02	.54
	High	12	2.09				
Organization	Medium	14	1.83	45.5	-1.98	.048	.39
	High	12	2.09				
Style & Quality of Expression	Low	7	1.40	17	-2.11	.04	.49
	High	12	1.65				
Mechanics	Low	7	0.79	15	-2.54	.01	.55
	Medium	14	0.91				

Mann-Whitney *U* tests revealed no statistically significant differences for the other paired groups for component scores assigned to high-quality essays.

Non-parametric tests were also conducted to compare the mean scores assigned to rubric components across three self-described rater experience groups (Gr1, $n_{\text{self-described low-experienced}} = 7$; Gr2, $n_{\text{self-described medium-experienced}} = 14$; Gr3, $n_{\text{self-described high-experienced}} = 12$) for low-quality essays. A Kruskal-Wallis test revealed statistically significant differences in the mean scores assigned to the grammar, style & quality of expression, and mechanics components of the rubric across the three self-described groups for low-quality essays. No statistically significant differences were found in the mean scores assigned to the content and organization components of the rubric. Table 16 summarizes the findings of the Kruskall-Wallis tests for rubric components for low-quality essays.

Table 16

Kruskall-Wallis Test Results for Low-quality Essay Component Scores by Self-described Experience Groups

Component	**χ^2 (2, *n* = 33)**	***p***	***Self-described low (Mdn)***	***Self-described medium (Mdn)***	***Self-described high (Mdn)***
Grammar	6.15	.046	0.66	0.78	0.80
Content	2.75	>.05	1.11	1.32	1.42
Organization	3.34	>.05	1.00	1.14	1.26
Style & Quality of Expression	6.95	.03	0.84	1.02	0.95
Mechanics	6.01	.05	0.48	0.65	0.57

As presented in Table 16, self-described low-experienced raters tended to record lower median scores to each component of the rubric. With the exception of the style & quality of expression and the mechanics components, the median score recorded by the self-described high-experienced raters tended to be higher than the other two self-described experience groups. For the style & quality of expression and the mechanics components, the self-described medium-experienced group recorded the highest median score. These findings are consistent with other trends in the data suggesting that higher-experienced raters and raters with higher self-described experience tended to assign higher scores than lower-experienced raters and raters with lower self-described experience.

Follow-up Mann-Whitney *U* tests were carried out to compare the differences between self-described groups for the grammar, style & quality of expression, and mechanics components for low-quality essays, as Kruskall-Wallis tests revealed significant differences between groups for these components. The statistically significant differences discovered by the Mann-Whitney U tests are summarized in Table 17.

Table 17

Mann-Whitney U Test Results for Low-quality Essay Component Scores by Self-described Experience Groups

Component	**Groups**	***n***	***Mdn***	***U***	***z***	***p***	***r***
Grammar	Low	7	0.66	17	-2.39	.02	.52
	Medium	14	0.78				
Grammar	Low	7	0.66	18	-2.03	.04	.47
	High	12	0.80				
Style & Quality of Expression	Low	7	0.84	15	-2.54	.01	.55
	Medium	14	1.02				
Mechanics	Low	7	0.48	17.5	-2.35	.02	.51
	Medium	14	0.65				

Mann-Whitney *U* tests revealed no statistically significant differences for the other paired groups for component scores assigned to low-quality essays.

In conclusion, self-described rater groups did not differ significantly in their total scores assigned to low-quality essays unlike the findings obtained from the comparison between reported experience groups. However, significant differences were derived from the total scores assigned to high-quality essays by self-described experience group. These findings are also contradictory to the findings obtained from the comparison between reported experience groups, which revealed no statistically significant differences in the scores assigned to high-quality essays. When further analyses were conducted on the individual rubric component scores, the only significant differences across experience groups were found for the scores assigned to the mechanics component for low-quality essays. However, between self-described experience groups, statistically significant differences were found for all rubric components for high-quality essays and the scores assigned to three rubric components (grammar, style and quality of expression, and mechanics) for low-quality essays. These results are striking in that they seem to suggest that self-described experience has a greater effect on the scores assigned to EFL essays than actual rating experience in terms of years.

Results for RQ3. The third research question is: What are the sources of score variation that contribute most (relatively) to the score variability of the analytic scores of EFL essays?

In order to determine the variance sources contributing to the analytic scoring, the person-by-rater-by-quality (*p x r x q*) random effects G-study was conducted. Table 18 illustrates the variance components and their relative contribution to the score variability.

Table 18

Variance Components for Random Effects P X R X Q Design

Variance Source	df	σ^2	%
P	24	2.59	45.3
R	32	0.09	1.6
Q	1	-0.03	0
PR	768	0.43	7.6
PQ	24	0.01	0.2
RQ	32	0.80	14
PRQ	768	1.79	31.3
Total	1649		100

Table 18 reveals that the largest variance component (45.3%) was due to persons, indicating that students differed in their writing performance as measured by the writing task. This result is desirable since the purpose of an assessment task is to differentiate students' writing abilities. The second greatest variance was attributable to the residual (31.3%), which was obtained from the interaction of raters, compositions, essay quality, and other unexplained unsystematic and systematic sources of errors. The third largest variance contributed to the score variability (14%) was the interaction between raters and essay quality, indicating that raters differed substantially while scoring compositions of distinct qualities. Table 18 shows that the fourth largest variance source was the interaction between persons and raters (7.6 of the total variance); this result means there was inconsistency between certain raters in terms of their judgements while assessing some certain essays. The remaining variance sources, including rater, essay quality, and person-by-quality, were negligible since their relative contributions to the variability of scores were small (1.6%, 0%, 0.2%, respectively).

In order to determine the variance sources contributing to the ratings assigned to high-quality essays, the person-by-rater (*p x r)* random effects G-study was conducted. Table 19 illustrates the variance components and their relative contribution to the score variability of high-quality essays.

Table 19

Variance Components for Random Effects P X R Design (High-quality Essays)

Variance Source	df	σ^2	%
P	24	0.20	6.8
R	32	1.15	39
PR	768	1.60	54.2
Total	824		100

According to Table 19, the greatest variance component was found to be the residual (54.2%), indicating that a large variance source is not explained in this design due to the interaction between persons, raters, and other systematic and unsystematic error sources. The second largest variance component followed by the residual was the rater facet (39%), indicating that raters' scores assigned to high-quality papers were markedly inconsistent. The smallest portion of variance was attributable to persons (6.8%), indicating that students did not differ significantly in their writing abilities. Although a larger variance percentage is desired from the object of measurement (persons), the relative contribution of the students in this design was considerably small. This can be seen as the result of homogeneous distribution of the students due to the particular selection of high-quality essays for the purpose of the analysis.

As for the determination of the variance sources contributing to the ratings assigned to low-quality essays, the person-by-rater (*p x r)* random effects G-study was conducted as well. Table 20 illustrates the variance components and their relative contribution to the score variability of low-quality essays.

Table 20

Variance Components for Random Effects P X R Design (Low-quality Essays)

Variance Source	df	σ^2	%
P	24	1.15	30.4
R	32	0.88	23.4
PR	768	1.74	46.2
Total	824		100

The results in Table 20 shows that the largest variance component was the residual (46.2%) because of the interaction between persons, raters, and other systematic and unsystematic error sources. The second largest variance component was persons followed by the residual (30%), indicating that students who were considered weak in their writing abilities performed differently in the given writing task. The remainder and smallest variance component was raters (23.4 %), which means that raters differed substantially in their scoring procedures while assessing low-quality papers.

To put it together, the writing task was effective to differentiate students in their writing abilities and raters were consistent while grading the essays collectively. However, while grading high-quality essays, raters differed substantially more in terms of leniency and severity compared to their ratings assigned to low-quality papers (39% and 23.4%, respectively). This might be related to the students' writing abilities within each essay quality group in that students performed more differently in low-quality essays (30.4%) compared to the writing performance exhibited in high-quality essays (6.8%). These results indicate that different interaction patterns occurred between persons, raters, and essay quality, indicating that low-quality essays were scored more similarly while raters applied more various scoring standards to assess high-quality essays.

Calculation of generalizability and dependability coefficients. Using the person-by-rater-by-quality (*p x r x q*) random effects G-study design for all papers, and person-by-rater (*p x r*) random effects design for high-quality and low-quality essays individually, the dependability coefficient (denoted as Φ) and generalizability coefficient (denoted as Ep^2 or G) were calculated. The results are presented in Table 21.

Table 21

Generalizability and Dependability Coefficients for Essay Ratings

Essays	N_{Essays}	N_{Rates}	Ep^2	Φ
All essays	50	33	.98	.98
High-quality	25	33	.81	.71
Low-quality	25	33	.96	.94

As shown in Table 21, the generalizability and dependability coefficients obtained for all papers with the current 50 essays and 33-rater scenario (Ep^2 = .98 and Φ = .98) were higher than those of obtained for low-quality essays (.96 and .94, respectively) and high-quality essays (.81 and .71, respectively). The results showed that while Ep^2 and Φ coefficients were obtained for the low-quality essays as well as all essays including both qualities were higher, the G-study analysis yielded lower Ep^2 and Φ coefficients for high-quality essays.

Followed by the calculation of G-coefficients and dependability indices, various D-studies based on completely crossed designs for three sets of essays (i.e. *p x r x q* for all essays and *p x r* for high-quality and low-quality essays) were conducted. The purpose of this procedure is to estimate the most suitable scoring design in similar assessment contexts. While designing D-studies, it is assumed that increasing the number of facets in a G-study design will produce higher Ep^2 and Φ coefficients. However, given that high coefficients were obtained for all essays and low-quality essays, a scenario in which the number of raters were decreased was planned until the Φ indices were achieved at acceptable level (i.e., above .80). However, the number of raters were increased in the scenario designed for high-quality essays since the Ep^2 and Φ coefficients were lower in the current scenario. Table 22 illustrates the generalizability and dependability coefficients in different scenarios in which the number of raters are manipulated.

Table 22

Generalizability and Dependability Coefficients for All, High-, and Low-quality Essays

All essays (N = 50, $p \times r \times q$)	N_{Raters}	Ep^2	Φ
	33	.98	.98
	23	.98	.97
	13	.96	.95
	3	**.85**	**.81**
Low-quality essays (n = 25, $p \times r$)	N_{Raters}	Ep^2	Φ
	33	.96	.94
	23	.94	.91
	13	.90	.85
	10	**.87**	**.81**
High-quality essays (n = 25, $p \times r$)	N_{Raters}	Ep^2	Φ
	33	.81	.71
	48	.86	.78
	53	.87	.79
	58	**.88**	**.81**

Table 22 shows that using the current G-study design, acceptable Ep^2 and Φ coefficients (i.e. above .80) would be obtained when the number of raters decreased down to three for all essays combining both high- and low-qualities. As for low-quality essays, acceptable Ep^2 and Φ coefficients would be obtained if a total of 10 raters were included. However, for high-quality essays, the desired generalizability and dependability coefficients can be obtained only if the number of raters are increased up to 58. An extended list of Ep^2 and Φ coefficients for the aforementioned analyses can be found in the appendix page (see Appendix Q).

Results for RQ4. The fourth research question is: Does the reliability (e.g., dependability coefficients for criterion-referenced score interpretations and generalizability coefficients for norm-referenced scores interpretations) of the analytic scores of raters differ based on their amount of experience?

In order to answer the fourth research question, G-study tests were conducted to compare the degree of agreement between raters within each experience group for their scores assigned to the essays using the person-by-rater-by-quality *(p x r x q)* random effects G-study design for all papers. Table 23 illustrates the inter-rater reliability coefficients obtained from the scores assigned to all essays that included both low-quality and high-quality compositions.

Table 23

Generalizability and Dependability Coefficients for All Essay Scores

Rater Group	N_{Raters}	N_{Essays} (50)	Ep^2	Φ
Low-experienced	13		.95	.93
Medium-experienced	10	Mixed quality	.93	.92
High-experienced	10		.95	.93

As can be seen in Table 23, high Ep^2 and Φ indices were obtained from the essay scores assigned by each of the experience rater group, indicating a perfect degree of agreement between raters within each experience group. Using person-by-rater random effects design *(p x r)* for high-quality and low-quality essays individually, the same tests were conducted to compare Ep^2 and Φ coefficients for low- and high-quality essays between raters within each experience group. Table 24 illustrates the coefficients for low-quality papers.

Table 24

Generalizability and Dependability Coefficients for Low-quality Essay Scores

Rater Group	N_{Raters}	N_{Essays} (25)	Ep^2	Φ
Low-experienced	13	Low-quality	.89	.85
Medium-experienced	10		.85	.79
High-experienced	10		.87	.85

As can be seen in Table 24, analytic scoring of the low-quality essays resulted in high generalizability and dependability coefficients for all experience groups. While low-experienced raters' scoring yielded the highest G-coefficient, slightly lower G-coefficients were obtained from high- and medium-experienced raters' scores assigned to low-quality essays (.87 and .85, respectively). As for dependability coefficients, the ratings of low-experienced and high-experienced rater groups produced higher Φ index (.85) compared to the coefficient obtained from the medium-experienced raters' scoring (Φ = .79). Using person-by-rater random effects design *(p x r)* G-studies were conducted for the scorings of rater experience groups for high-quality essays. Table 25 displays G- and dependability coefficients calculated for the analytic scoring of high-quality essays.

Table 25

Generalizability and Dependability Coefficients for High-quality Essay Scores

Rater Group	N_{Raters}	N_{Essays} (25)	Ep^2	Φ
Low-experienced	13	High-quality	.57	.44
Medium-experienced	10		.52	.46
High-experienced	10		.47	.27

Although high Ep^2 and Φ coefficients were obtained from the scores of mixed-quality essays and low-quality essays, significantly lower G- and dependability coefficients were observed for high-quality essays, indicating differences between raters within each group in terms rating leniency and severity. The high-experienced raters' scoring resulted in the lowest Ep^2 and Φ indices, implying that the raters with the highest level of scoring experience were the most inconsistent group while scoring high-quality compositions.

Followed by the calculation of G- and dependability coefficients, various D-studies based on completely crossed designs for three sets of essays (i.e. *p x r x q* for all essays and *p x r* for high-quality and low-quality essays) were conducted for each rater experience group. Given that higher coefficients were obtained for all essays and low-quality essays, a scenario in which the number of raters were decreased was planned until the Φ indices were achieved at acceptable level (i.e., above .80). However, the number of raters were increased in the scenario designed for high-quality essays since the Ep^2 and Φ coefficients were lower in the current scenario. Table 26 illustrates the G- and dependability coefficients obtained from decision studies when the number of raters are manipulated.

Table 26

Generalizability and Dependability Coefficients for Low-experienced Raters

All essays (*N* = 50)	N_{Raters}	Ep^2	Φ
	13	.95	.93
	12	.95	.92
	6	.91	.86
	5	**.89**	**.83**
Low-quality essays (*n* = 25)	N_{Raters}	Ep^2	Φ
	13	.89	.85
	12	.89	.84
	10	**.87**	**.81**
	21	.93	.90
High-quality essays (*n* = 25)	N_{Raters}	Ep^2	Φ
	13	.57	.44
	28	.74	.62
	48	.83	.74
	73	**.88**	**.81**

As can be seen in Table 26, acceptable dependability coefficients (i.e. above .80) would be obtained for high- and low-quality mixed essays if the number of raters was decreased to five (Φ = .83). As for the scorings of low-quality essays, using 10 raters would yield an acceptable dependability coefficient (Φ = .83). However, increasing the number of raters from 13 to 73 would elevate the dependability coefficient from .44 to .81 for high-quality essays. An extended list of Ep^2 and Φ coefficients for the aforementioned analyses conducted on the ratings of low-experienced raters can be found in the appendix page (see Appendix R). In the same manner, Table 27 shows the fluctuation in G- and dependability coefficients pertaining to the scoring of the medium-experienced rater group.

Table 27

Generalizability and Dependability Coefficients for Medium-experienced Raters

All essays (*N* = 50)	N_{Raters}	Ep^2	Φ
	10	.93	.92
	6	.89	.87
	5	.87	.84
	4	**.84**	**.81**
Low-quality essays (*n* = 25)	N_{Raters}	Ep^2	Φ
	10	.85	.79
	11	**.86**	**.81**
	15	.89	.85
	23	.93	.90
High-quality essays (*n* = 25)	N_{Raters}	Ep^2	Φ
	10	.52	.46
	25	.73	.68
	35	.79	.75
	50	**.84**	**.81**

According to Table 27, using four raters for assessing the essay set combining distinct qualities would be enough to obtain an acceptable dependability coefficient (.81). With respect to low-quality essays, including one more rater in the current scenario would increase the Φ index from .79 to .81. Nonetheless, a minimum of 50 raters would be needed to obtain an acceptable dependability coefficient (Φ = .81) from the scoring of high-quality compositions.

An extended list of Ep^2 and Φ coefficients for the aforementioned analyses conducted on the ratings of medium-experienced raters can be found in the appendix page (see Appendix S).

Finally, G- and dependability coefficients for the scoring of the high-experienced rater group were calculated. Table 28 shows the change in coefficients indices when the rater facet is manipulated for the three sets of essays.

Table 28

Generalizability and Dependability Coefficients for High-experienced Raters

All essays (*N* = 50)	N_{Raters}	Ep^2	Φ
	10	.95	.93
	6	.91	.89
	4	**.88**	**.84**
	3	.84	.80
Low-quality essays (*n* = 25)	N_{Raters}	Ep^2	Φ
	10	.87	.85
	9	.86	.83
	8	**.85**	**.82**
	7	.83	.80
High-quality essays (*n* = 25)	N_{Raters}	Ep^2	Φ
	10	.47	.27
	60	.80	.69
	110	.91	.80
	115	**.91**	**.81**

Table 28 reveals that decreasing the number of raters to four for all essays and to eight for low-quality essays would still produce acceptable dependability coefficients (.84 and .82, respectively). However, 115 raters would be needed to be involved in the scoring process of high-quality essays to obtain a dependability index of .81. An extended list of Ep^2 and Φ coefficients for the aforementioned analyses conducted on the ratings of high-experienced raters can be found in the appendix page (see Appendix T).

As a result, using the raters in an economic way by decreasing their number would be feasible for the essays of distinct qualities together and low-quality essays, except for the medium-experienced rater group’s scoring, in which one more rater is necessary for a dependability index above .80 for low-quality essays. However, none of the scenarios is realistic for the scoring of high-quality essays, as a large number of raters would be necessary to achieve the desired dependability coefficients.

Qualitative Data Analysis Results

This section includes the analysis of two sets of data: think-aloud protocols recorded by raters while scoring the essays, and written score explanations recorded by raters for each essay. The primary data set includes the TAPs that were obtained from 28 raters (five participants failed to conduct TAPs during their assessments). Each rater conducted 16 TAPs in different audio files. While eight of the essays for which TAPs were conducted were labeled as high-quality essays, the remaining eight essays were considered low-quality. Two of the raters failed to record their thoughts for one essay each, eliminating one high-quality and one low-quality essay from the TAPs data set. Thus, the researcher was able to obtain a total number of 446 TAPs for qualitative data analysis. The TAP recordings ranged from 2:17 minutes to 23:21 minutes in length, with an average recoding time of 8:27 minutes. To analyze the data, a coding scheme adapted from Cumming et al. (2002) was used. The scheme was comprised of 35 decision-making behaviors, which were divided into two categories, labelled as interpretation and judgements strategies, and categorized along three foci, including self-monitoring focus, rhetorical and ideational focus, and language focus. In light of the aforementioned explanations, the presentation of the qualitative data follows a cumulative approach while considering essay quality and the experience level of the rater groups.

Following the coding procedure, a total number of 14,562 decision-making strategies were identified. While raters employed 7,827 decision-making strategies for low-quality essays, 6,735 strategies were used for high-quality essays. A total of 223 TAPs were obtained for each essay quality, and a number of 30.2 strategies were used on average per high-quality essay, while an average number of 35.1 strategies were used per low-quality essay, resulting in an average number of 32.65 strategies per essay in general. These trends suggest that raters relied on more strategies while grading low-quality essays compared to high-quality essays. However, while further interpreting the data by rater experience groups, the numbers were converted to percentages because: a) the number of raters was uneven across each experience group and b) the numbers might reflect “verbosity and personal rhythms” of the raters (Cumming et al., 2002, p. 78) while percentages diminish these effects.

In addition to recording think-aloud protocols, raters were asked to provide three written explanations for their assigned score to each essay. A space was provided at the beginning of each essay sheet for raters to write their three reasons for the assigned score, and raters were asked to provide this information for each essay after assigning a score and before beginning their assessment of the next essay. While a few raters failed to provide three reasons for every essay, other raters provided more than three reasons to some essays. As such, the total number of reasons identified exceeded the researcher’s expected count of 4,950 reasons (33 raters x 50 essays x 3 reasons per essay). Instead, the total number of reasons provided by the raters was 5,425, or approximately 3.3 reasons per paper.

The reasons provided by the raters were coded inductively, with the researcher identifying the theme and connotation (positive or negative) of each explanation. A total of 24 themes were identified. These themes were then reduced to 18 themes for the purposes of data presentation and analysis. The following guidelines served to clarify the major themes:

- Grammar—includes reasons relating to syntax, morphology, and references to grammatical mistakes in the essay.
- Content—includes an assessment of the ideas or level of critical thinking evident in the essay; it also includes reasons directly related to the quality of the essay's content.
- Language overall—includes reasons that refer to the general quality of the language used in the essay.
- Mechanics—includes reasons commenting on aspects of spelling, punctuation, capitalization, or an overall assessment of mechanics.
- Register—includes reasons related to the style and quality of expression in the essay. It also includes reasons that identify direct translation from the L1 in the essay. The researcher chose to include these comments under the theme of register because L1 translation is listed under the style component of the analytic rubric used to assess the essays.
- Raters were found to frequently refer to the quality of the introduction and conclusion paragraphs in their explanations. For this reason, "introduction" and "conclusion" were created as separate themes, distinct from organization, and include reasons that explicitly comment on the quality of the introduction or conclusion respectively.
- The category of "other" includes the themes of title, length, and layout, which were identified but rarely used by the raters. In total, title was mentioned in eight reasons, and length and layout were each mentioned once.

The data collected from the written score explanations provided by raters were used to triangulate the findings from the think-aloud protocols in answering the two qualitative research questions, **RQ5** and **RQ6**, the findings for which are presented below.

Results for RQ5. The fifth research question is: How do raters make decisions while rating different quality EFL essays analytically?

The first qualitative research question (RQ5) enquires into how raters make decisions while rating different quality EFL essays analytically. As mentioned in the previous section, raters tended to rely on more strategies while grading low-quality essays compared to high-quality essays, with an average of 35.1 strategies for low-quality and 30.2 strategies for high-quality essays. The analyses and tables presented in this section aim to explore the strategies used for each essay quality in more detail by describing the distribution of strategy use by raters.

In order to examine the distribution of decision-making strategies in terms of strategy type—interpretation or judgment—descriptive statistical analysis was conducted. Figure 15 shows the total number of strategies employed by raters and their distribution based on essay quality and strategy type.

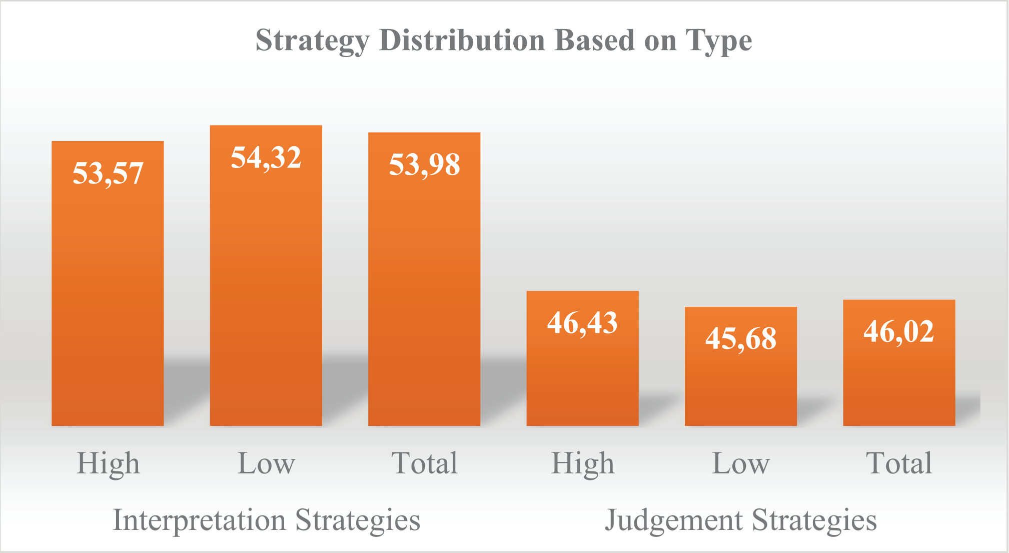


Figure 15. Distribution of strategy type based on essay quality.

While 53.98% of the total strategies recorded were interpretation strategies, 46.02% of them were considered judgement strategies. When considering essay quality, there is not a large difference to be observed between strategy types. As can be seen in Figure 15, low-quality essays attained slightly more interpretation strategies while raters used slightly more judgement strategies for high-quality essays.

Next, the strategies used were divided into three categories regarding their focus: a) self-monitoring, b) rhetorical and ideational, and c) language. Figure 16 displays the distribution of the strategies based on focus and essay quality.

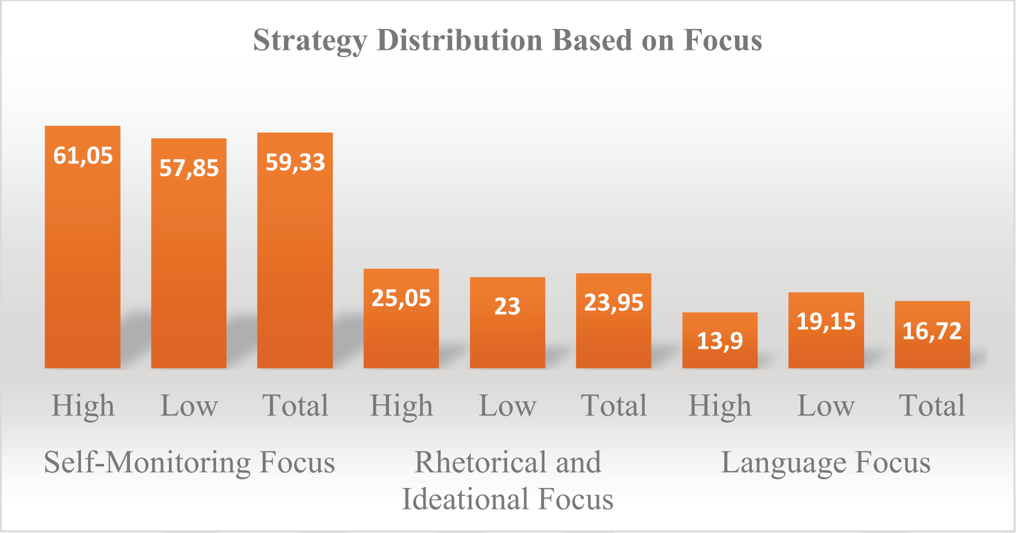


Figure 16. Distribution of strategies based on focus and essay quality.

According to Figure 16, the most commonly used strategies belonged to the self-monitoring focus (59.33%) followed by the rhetorical and ideational focus (23.95%) and language focus (16.72%), respectively. The same trend can be observed both for high-quality and low-quality essays. However, while raters used slightly more strategies in self-monitoring and rhetorical and ideational foci for high-quality essays, low-quality essays attracted more language related strategies compared to high-quality essays.

After examining strategy use in terms of type and focus, the researcher sought to compare the use of individual strategies by raters across experience groups and essay qualities. Table 29 describes the most commonly used decision-making behaviors by all raters for all essays. It rank orders the 35 decision-making behaviors included in the coding scheme. While

Table 29 refers to strategy use for all essays, Table 30 compares the frequency of strategies used for low and high-quality essays.

Table 29

The Most Commonly Used Decision-making Behaviors by All Raters for All Essays

Decision-Making Behavior	**%**
Read or reread text	**27.98**
Articulate or revise scoring	**11.91**
Read or interpret scoring scale	**11.41**
Summarize ideas or propositions	**6.10**
Consider syntax or morphology	**4.13**
Articulate general impression	**3.23**
Assess task completion or relevance	**3.21**
Edit phrases for interpretation	**2.75**
Consider spelling or punctuation	**2.72**
Assess reasoning, logic, or topic development	**2.70**
Assess coherence	2.19
Assess text organization	2.05
Rate ideas or rhetoric	2.05
Assess style, register, or genre	1.90
Discern rhetorical structure	1.63
Consider or personal response, expectations or biases	1.45
Consider lexis	1.36
Assess comprehensibility	1.35
Interpret ambiguous or unclear phrases	1.30
Assess quantity of written production	1.22
Summarize, distinguish, or tally judgements collectively	1.21
Classify errors into types	1.07
Envision personal situation of the writer	0.73
Rate language overall	0.62
Identify redundancies	0.61
Scan or skim text	0.57
Consider error frequency	0.55
Decide on macro-strategy for reading and rating	0.38
Assess fluency	0.37
Consider gravity of errors	0.34
Compare with other compositions or "anchors"	0.24
Observe layout	0.23
Assess interest, originality, or creativity	0.20
Read or interpret essay prompt	0.19
Define or revise own criteria	0.03
Total	100

Note. The percentage values have been rounded to the nearest hundredth.

Table 30

The Most Commonly Used Decision-making Behaviors for High- and Low-quality Essays

Low-quality Essays		**High-quality Essays**	
Decision-Making Behaviors	**%**	**Decision-Making Behaviors**	**%**
Read or reread text	29.27	Read or reread text	26.49
Articulate or revise scoring	11.05	Articulate or revise scoring	12.90
Read or interpret scoring scale	10.51	Read or interpret scoring scale	12.44
Summarize ideas or propositions	4.93	Summarize ideas or propositions	7.45
Consider syntax or morphology	4.69	Articulate general impression	4.38
Consider spelling or punctuation	3.82	Consider syntax or morphology	3.49
Assess task completion or relevance	3.32	Assess task completion or relevance	3.09
Edit phrases for interpretation	2.94	Assess reasoning, logic, or topic development	3.06
Assess style, register, or genre	2.61	Rate ideas or rhetoric	2.81
Assess reasoning, logic, or topic development	2.39	Edit phrases for interpretation	2.52
Articulate general impression	2.24	Assess coherence	2.21
Assess coherence	2.17	Discern rhetorical structure	1.95
Assess text organization	2.15	Assess text organization	1.95
Assess comprehensibility	2.06	Consider lexis	1.75
Interpret ambiguous or unclear phrases	1.85	Assess quantity	1.46
Consider own personal response, expectations or biases	1.57	Consider spelling or punctuation	1.44
Classify errors into types	1.44	Summarize judgements collectively	1.40
Rate ideas or rhetoric	1.39	Consider own personal response, expectations or biases	1.31
Discern rhetorical structure	1.37	Assess style, register, or genre	1.07
Summarize judgements collectively	1.05	Assess fluency	0.68
Envision personal situation of writer	1.03	Interpret ambiguous or unclear phrases	0.67
Assess quantity	1.02	Rate language overall	0.65
Consider lexis	1.02	Classify errors into types	0.64
Consider error frequency	0.80	Scan or skim text	0.59
Identify redundancies	0.64	Identify redundancies	0.58
Rate language overall	0.60	Assess comprehensibility	0.52
Scan or skim text	0.55	Decide on macro-strategy	0.49
Consider gravity of errors	0.40	Envision personal situation of writer	0.39
Decide on macro-strategy	0.28	Compare with other compositions	0.39
Observe layout	0.26	Consider gravity of errors	0.28
Assess originality	0.18	Consider error frequency	0.25
Read prompt	0.17	Read prompt	0.22
Compare with other compositions	0.11	Assess originality	0.22
Assess fluency	0.10	Observe layout	0.21
Define or revise own criteria	0.01	Define or revise own criteria	0.06
Total	100	**Total**	100

Note. The percentage values have been rounded to the nearest hundredth.

As can be seen from Tables 29 and 30, the most commonly used strategies for all essays were, “read or reread text,” “articulate or revise scoring,” “read or interpret scoring scale,” and “summarize ideas or proposition.” These four strategies accounted for 57.40% of strategy use for all essays, 59.29% of strategy use for high-quality essays, and 55.77% of strategy use for low-quality essays. Thus, these top four strategies were more commonly used for high-quality essays compared to low-quality essays. When considering the remaining six strategies found in the top ten for each essay quality, it can be seen that the strategies, “consider syntax and morphology” and “consider spelling or punctuation,” were more commonly used for low-quality essays compared to high-quality essays. The strategy, “assess style, register, or genre,” was also more commonly used for low-quality essays and did not appear in the top ten most common strategies for high-quality essays. Rather, “assess style, register, or genre” ranked 19th for high-quality essays, although it ranked ninth for low-quality essays.

Similarly, the strategies “articulate general impression” and “rate ideas or rhetoric” were more commonly used for high-quality essays than low-quality essays. Both strategies appeared in the top ten for high-quality essays but were ranked 11th and 18th respectively for low-quality essays. Considered collectively, these trends suggest that raters focused more on style, grammar, and mechanics when rating low-quality essays but more on ideas, rhetoric, and their general impression of the essay when rating high-quality essays.

To triangulate these findings, qualitative data collected from the written explanations provided for the scoring of each essay were analyzed. Table 31 presents the most commonly given reasons for the scores assigned to high-quality essays, including overall frequency and percentage as well as the breakdown in terms of positive or negative connotation.

Table 31

The Most Commonly Given Reasons for High-quality Essays

Theme	**Overall Reasons**		**Positive Reasons**		**Negative Reasons**	
	f	**%**	***f***	**%**	***f***	**%**
Organization	463	17.11	378	13.97	85	3.14
Grammar	402	14.86	317	11.71	85	3.14
Topic development	290	10.72	197	7.28	93	3.44
Relevance	249	9.20	127	4.69	122	4.51
Lexis	194	7.17	154	5.69	40	1.48
Content	162	5.99	140	5.17	22	0.81
Language overall	161	5.95	141	5.21	20	0.74
Coherence	142	5.25	65	2.40	77	2.85
Thesis	127	4.69	85	3.14	42	1.55
Mechanics	107	3.95	62	2.29	45	1.66
Register	98	3.62	34	1.26	64	2.37
Comprehensibility	94	3.47	82	3.03	12	0.44
Introduction	79	2.92	64	2.37	15	0.55
Fluency	60	2.22	38	1.40	22	0.81
Conclusion	32	1.18	7	0.26	25	0.92
Overall impression	22	0.81	21	0.78	1	0.04
Redundancy	20	0.74	0	0.00	20	0.74
Other	4	0.15	0	0.00	4	0.15
Total	**2706**	**100**	**1912**	**70.66**	**794**	**29.34**

Note. The percentage values have been rounded to the nearest hundredth.

A total of 2,706 reasons were identified for high-quality essays, the overwhelming majority of which were positive (70.66%). Organization and grammar were the two most commonly given reasons, followed by topic development, relevance, and lexis. These findings corroborate the data collected from think-aloud protocols in that syntax and morphology, task completion and relevance, and topic development were found to be common strategies used to assess high-quality essays. However, data derived from the three reasons found a much strong emphasize on organization and lexis than emerged from the decision-making behaviors used by raters to grade high-quality essays.

A similar analysis was conducted on the score explanations provided for low-quality essays. Table 32 summarizes the mostly commonly identified reasons provided for the scoring of low-quality essays.

Table 32

The Most Commonly Given Reasons for Low-quality Essays

Theme	**Overall Reasons**		**Positive Reasons**		**Negative Reasons**	
	f	**%**	***f***	**%**	***f***	**%**
Grammar	501	18.43	87	3.20	414	15.23
Organization	341	12.54	93	3.42	248	9.12
Mechanics	290	10.67	15	0.55	275	10.11
Relevance	248	9.12	57	2.10	191	7.02
Topic development	195	7.17	46	1.69	149	5.48
Coherence	144	5.30	16	0.59	128	4.71
Language overall	144	5.30	37	1.36	107	3.94
Thesis statement	141	5.19	30	1.10	111	4.08
Register	134	4.93	11	0.40	123	4.52
Comprehensibility	130	4.78	29	1.07	101	3.71
Lexis	118	4.34	43	1.58	75	2.76
Content	113	4.16	38	1.40	75	2.76
Introduction	71	2.61	20	0.74	51	1.88
Fluency	47	1.73	9	0.33	38	1.40
Redundancy	46	1.69	1	0.04	45	1.66
Conclusion	31	1.14	5	0.18	26	0.96
Overall impression	19	0.70	7	0.26	12	0.44
Other	6	0.22	3	0.11	3	0.11
Total	**2719**	**100**	**547**	**20.12**	**2172**	**79.88**

Note. The percentage values have been rounded to the nearest hundredth.

A total of 2,719 reasons were identified for low-quality essays, almost identical to the number of reasons provided for high-quality essays. As they tended to provide positive reasons for high-quality essays, raters tended to provide negative reasons for their scoring of low-quality essays, although they were more negative in their assessment of low-quality essays than they were positive in their assessment of high-quality essays: almost 80% of reasons for low-quality essays were negative, whereas about 70% of reasons for high-quality essays were positive.

The most common reasons identified for the scoring of low-quality essays were grammar, organization, and mechanics. While 10.67% of reasons for low-quality essays related to mechanics, this category accounted for less than 4% of reasons for high-quality essays. The difference in attention to mechanics for high- and low-quality essays supports the findings of the think-aloud protocols, in which it was found that raters tended to focus more on elements of language such as grammar, spelling, and punctuation when assessing low-quality essays than high-quality essays. Similarly, while 5.99% and 10.72% of reasons related to content and topic development respectively for high-quality essays, these two themes accounted for only 4.16% and for 7.17%, respectively, of the reasons provided for low-quality essays, corroborating the findings from the think-aloud protocol data that suggest that raters tend to focus more on ideas when assessing high-quality essays.

Results for RQ6. The sixth research question is: How is rating experience related to EFL raters' decision-making processes and the aspects of writing they attend to?

The second qualitative research question asked how rating experience related to EFL raters' decision-making processes and the aspects of writing that they attend to while rating EFL compositions. To answer this question, data collected from TAPs were analyzed to compare the decision-making behaviors of low-, medium-, and high-experienced raters when scoring essays. To begin, Figure 17 summarizes the distribution of strategies based on type by experience groups.

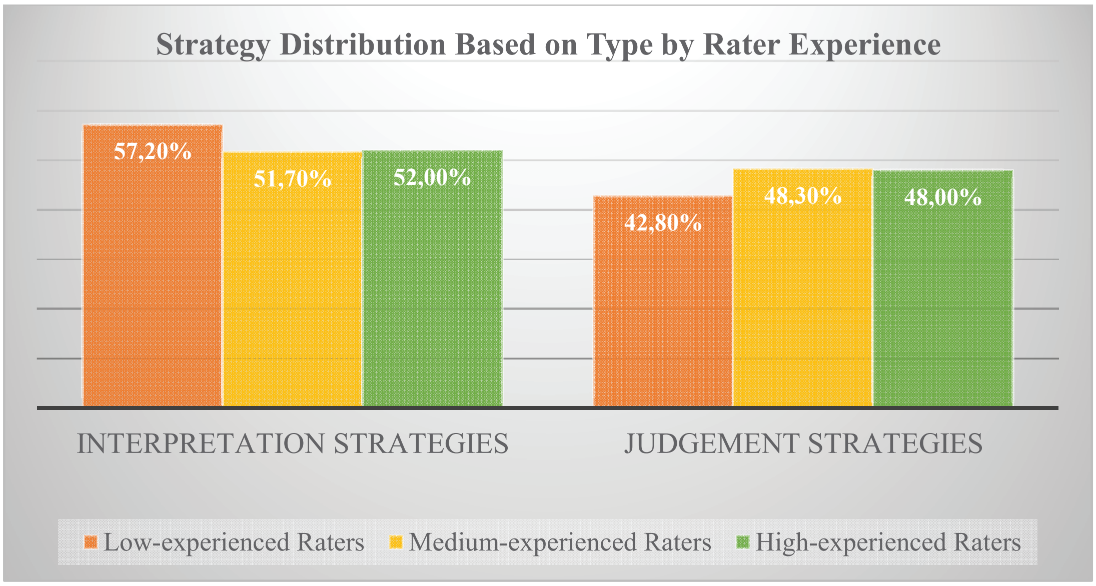


Figure 17. Strategy distribution based on type by rater experience.

According to Figure 17, low-experienced raters used more interpretation strategies (57.20%) than their more experienced peers whereas medium- and high-experienced raters employed judgment strategies more frequently than raters with less experience did. These results indicate that raters with less experience dealt with interpreting the text features to provide a basis for their judgements. When considering the strategy focus, Figure 18 illustrates the distribution of strategies by focus according to rater experience groups.

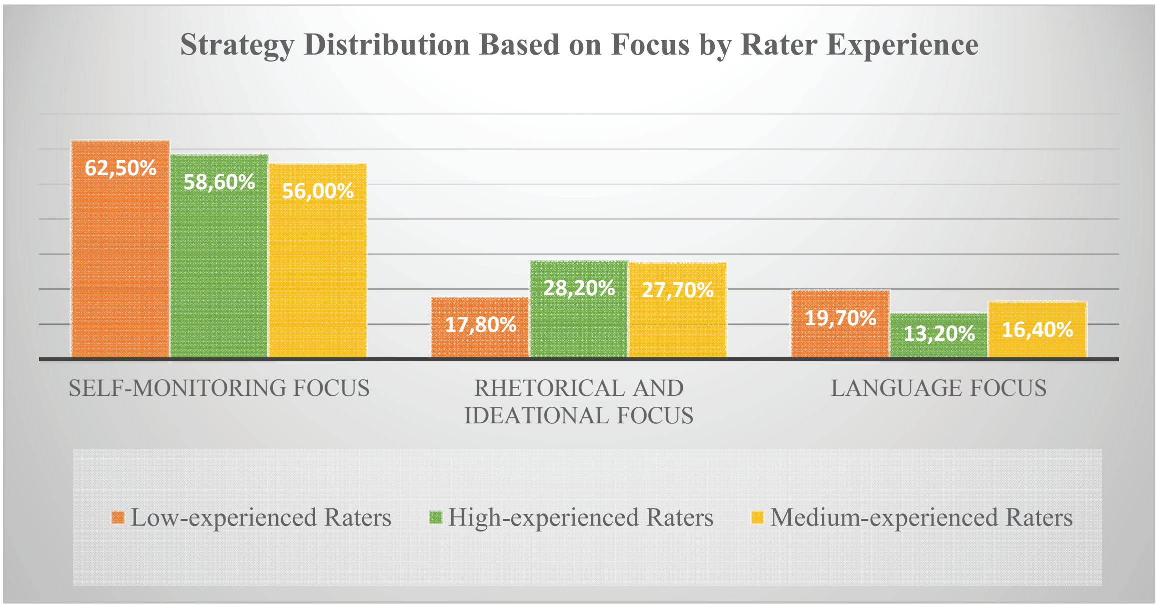


Figure 18. Strategy distribution based on focus by rater experience.

As can be seen in Figure 18, low-experienced raters focused more on language (19.70%) than medium- (16.40%) or high- (13.20%) experienced raters. Added to that, raters with less experience tended to use strategies pertaining to the self-monitoring focus more frequently than more experienced raters. However, medium- and high-experienced raters attended to decision-making behaviors related to rhetorical and ideational focus more than the low-experienced raters did.

Table 33 compares the distribution of decision-making behaviors for each of the 35 strategies included in the coding schema across the three experience groups, while Table 34 and Table 35 present data for the same comparison for low-quality and high-quality essays respectively.

Table 33

Distribution of Decision-making Behaviors for All Essays Based on Raters' Experience Groups

	Rater experience group		
	Low	**Medium**	**High**
Decision-Making Behaviors	**%**	**%**	**%**
Read or interpret prompt	0.12	0.35	0.13
Read or reread text	**30.44**	**27.11**	**25.69**
Envision personal situation of the writer	0.42	1.04	0.84
Scan or skim text	0.33	0.51	0.93
Read or interpret scoring scale	**12.67**	**9.17**	**11.95**
Decide on macro-strategy for reading and rating	0.16	0.65	0.40
Consider own personal response, expectations or biases	0.75	**2.55**	1.29
Define or revise own criteria	0.05	0.02	0.02
Compare with other compositions or "anchors"	0.19	0.23	0.31
Summarize, distinguish, or tally judgements collectively	1.59	1.11	0.82
Articulate general impression	**2.58**	**3.52**	**3.78**
Articulate or revise scoring	**13.17**	**9.70**	**12.41**
Interpret ambiguous or unclear phrases	1.41	1.67	0.82
Discern rhetorical structure	1.11	1.62	2.31
Summarize ideas or propositions	**4.32**	**6.81**	**7.68**
Assess reasoning, logic, or topic development	1.90	**3.52**	**2.93**
Assess task completion or relevance	**2.28**	**3.64**	**4.00**
Assess coherence	**2.09**	2.27	2.24
Assess interest, originality, or creativity	0.10	0.37	0.16
Identify redundancies	0.57	0.72	0.56
Assess text organization	1.18	**2.69**	**2.55**
Assess style, register, or genre	1.64	1.94	2.18
Rate ideas or rhetoric	1.18	2.43	**2.78**
Observe layout	0.30	0.25	0.13
Classify errors into types	1.29	1.07	0.80
Edit phrases for interpretation	**4.77**	2.15	0.73
Assess quantity of written production	0.98	0.88	1.87
Assess comprehensibility	1.50	1.50	1.00
Consider gravity of errors	0.28	0.58	0.20
Consider error frequency	0.51	0.44	0.71
Assess fluency	0.33	0.58	0.22
Consider lexis	1.41	1.46	1.20
Consider syntax or morphology	**4.16**	**4.86**	**3.40**
Consider spelling or punctuation	**3.90**	1.94	1.95
Rate language overall	0.31	0.65	1.00
Total	100	100	100

Note. The percentage values have been rounded to the nearest hundredth.

Table 34

Distribution of Decision-making Behaviors for Low-quality Essays Based on Raters' Experience Groups

	Rater experience group		
	Low	**Medium**	**High**
Decision-Making Behaviors	**%**	**%**	**%**
Read or interpret prompt	0.16	0.26	0.08
Read or reread text	**32.20**	**28.13**	**26.64**
Envision personal situation of the writer	0.68	1.29	1.24
Scan or skim text	0.36	0.26	1.08
Read or interpret scoring scale	**11.31**	**8.62**	**11.34**
Decide on macro-strategy for reading and rating	0.10	0.51	0.29
Consider own personal response, expectations or biases	0.75	**2.87**	1.37
Define or revise own criteria	0.03	0.00	0.00
Compare with other compositions or "anchors"	0.03	0.17	0.17
Summarize, distinguish, or tally judgements collectively	1.36	1.07	0.62
Articulate general impression	1.66	2.66	2.57
Articulate or revise scoring	**11.86**	**9.26**	**11.75**
Interpret ambiguous or unclear phrases	1.82	2.66	1.12
Discern rhetorical structure	0.81	1.24	2.19
Summarize ideas or propositions	**3.25**	**5.75**	**6.29**
Assess reasoning, logic, or topic development	1.56	**3.00**	**2.85**
Assess task completion or relevance	**2.57**	**3.52**	**4.10**
Assess coherence	2.05	2.19	2.32
Assess interest, originality, or creativity	0.13	0.30	0.12
Identify redundancies	0.55	0.73	0.66
Assess text organization	1.14	2.79	**2.81**
Assess style, register, or genre	**2.21**	**3.00**	**2.73**
Rate ideas or rhetoric	0.58	1.72	2.11
Observe layout	0.36	0.17	0.21
Classify errors into types	1.75	1.37	1.12
Edit phrases for interpretation	**5.23**	2.27	0.66
Assess quantity of written production	0.81	0.69	1.61
Assess comprehensibility	**2.21**	2.40	1.53
Consider gravity of errors	0.36	0.60	0.25
Consider error frequency	0.68	0.60	1.16
Assess fluency	0.06	0.09	0.17
Consider lexis	0.97	1.16	0.95
Consider syntax or morphology	**4.81**	**5.06**	**4.18**
Consider spelling or punctuation	**5.30**	**3.04**	**2.69**
Rate language overall	0.29	0.56	1.03
Total	100	100	100

Note. The percentage values have been rounded to the nearest hundredth.

Table 35

Distribution of Decision-making Behaviors for High-quality Essays Based on Raters' Experience Groups

	Rater experience group		
	Low	**Medium**	**High**
Decision-Making Behaviors	**%**	**%**	**%**
Read or interpret prompt	0.08	0.45	0.19
Read or reread text	**28.40**	**25.92**	**24.59**
Envision personal situation of the writer	0.11	0.75	0.38
Scan or skim text	0.30	0.81	0.77
Read or interpret scoring scale	**14.24**	**9.81**	**12.66**
Decide on macro-strategy for reading and rating	0.23	0.81	0.53
Consider own personal response, expectations or biases	0.75	2.16	1.20
Define or revise own criteria	0.08	0.05	0.05
Compare with other compositions or "anchors"	0.38	0.30	0.48
Summarize, distinguish, or tally judgements collectively	1.84	1.16	1.05
Articulate general impression	**3.64**	**4.53**	**5.18**
Articulate or revise scoring	**14.69**	**10.22**	**13.18**
Interpret ambiguous or unclear phrases	0.94	0.50	0.48
Discern rhetorical structure	1.47	2.06	**2.44**
Summarize ideas or propositions	**5.56**	**8.05**	**9.30**
Assess reasoning, logic, or topic development	**2.29**	**4.13**	**3.02**
Assess task completion or relevance	1.95	**3.77**	**3.88**
Assess coherence	**2.14**	2.37	2.16
Assess interest, originality, or creativity	0.08	0.45	0.19
Identify redundancies	0.60	0.70	0.43
Assess text organization	1.24	**2.57**	2.25
Assess style, register, or genre	0.98	0.70	1.53
Rate ideas or rhetoric	1.88	**3.27**	**3.55**
Observe layout	0.23	0.35	0.05
Classify errors into types	0.75	0.70	0.43
Edit phrases for interpretation	**4.24**	2.01	0.81
Assess quantity of written production	1.16	1.11	2.16
Assess comprehensibility	0.68	0.45	0.38
Consider gravity of errors	0.19	0.55	0.14
Consider error frequency	0.30	0.25	0.19
Assess fluency	0.64	1.16	0.29
Consider lexis	1.92	1.81	1.49
Consider syntax or morphology	**3.42**	**4.63**	**2.49**
Consider spelling or punctuation	**2.29**	0.65	1.10
Rate language overall	0.34	0.75	0.96
Total	100	100	100

Note. The percentage values have been rounded to the nearest hundredth.

In Table 35, the most common decision-making behaviors for each experience group are marked in bold. As can be seen, commonalities exist between the groups, particularly between the medium- and high-experienced groups. The most commonly used strategies with respect to language tend to concern grammar and mechanics, and the most commonly used strategies with respect to rhetoric and ideas tend to involve topic development, relevance, and coherence. Rearranging the data presented in Tables 33, 34 and 35, Table 36 compares the top ten most frequently used strategies for each experience group. The data in this table refers to rating behavior for all essays, combining both high and low-quality essays.

Table 36

The Most Common Decision-making Behaviors for All Essays

Low-experienced raters		**Medium-experienced raters**		**High-experienced raters**	
Decision-Making Behaviors	**%**	**Decision-Making Behaviors**	**%**	**Decision-Making Behaviors**	**%**
Read or reread text	30.44	Read or reread text	27.11	Read or reread text	25.69
Articulate or revise scoring	13.17	Articulate or revise scoring	9.70	Articulate or revise scoring	12.41
Read or interpret scoring scale	12.67	Read or interpret scoring scale	9.17	Read or interpret scoring scale	11.95
Edit phrases for interpretation	4.77	Summarize ideas or propositions	6.81	Summarize ideas or propositions	7.68
Summarize ideas or propositions	4.32	Consider syntax or morphology	4.86	Assess task completion or relevance	4.00
Consider syntax or morphology	4.16	Assess task completion or relevance	3.64	Articulate general impression	3.78
Consider spelling or punctuation	3.90	Articulate general impression	3.52	Consider syntax or morphology	3.40
Articulate general impression	2.58	Assess reasoning, logic, or topic development	3.52	Assess reasoning, logic, or topic development	2.93
Assess task completion or relevance	2.28	Assess text organization	2.69	Rate ideas or rhetoric	2.78
Assess coherence	2.09	Consider own personal response, expectations or biases	2.55	Assess text organization	2.55

Note. The percentage values have been rounded to the nearest hundredth.

As can be seen in Table 36, each of the three experience groups displayed the same top three decision-making behaviors: “read or reread text,” “articulate or revise scoring,” and “read or interpret scoring scale.” These three behaviors accounted for 56.27% of all strategies used by the low-experienced group, 45.98% of the strategies used by the medium-experienced group, and 50.06% of all strategy use by the high-experienced group. Across the three experience groups, seven of the top ten strategies used were the same. However, similarities in rating behaviors are more pronounced between medium- and high-experienced raters, who shared nine out of 10 behaviors. Medium- and high-experienced raters differed only in that medium-experienced raters recorded, “consider own personal response, expectations, or bias,” as the tenth most common strategy, while “rate ideas or rhetoric” appeared in the top ten strategies for high-experienced raters. Thus, medium- and high-experienced raters tended to display more similarities in their decision-making behaviors than low-experienced raters, who tended to put more emphasis on mechanics and focus more on language than their more experienced peers.

While Table 36 presented data for all essays, Tables 37 and 38 present rater behavior by experience group for each essay quality. First, Table 37 compares the top ten most frequently used strategies by each experience group when rating low-quality essays.

Table 37

The Most Common Decision-making Behaviors for Low-quality Essays

Low-experienced raters		**Medium-experienced raters**		**High-experienced raters**	
Decision-Making Behaviors	**%**	**Decision-Making Behaviors**	**%**	**Decision-Making Behaviors**	**%**
Read or reread text	32.20	Read or reread text	28.13	Read or reread text	26.64
Articulate or revise scoring	11.89	Articulate or revise scoring	9.26	Articulate or revise scoring	11.75
Read or interpret scoring scale	11.31	Read or interpret scoring scale	8.62	Read or interpret scoring scale	11.34
Consider spelling or punctuation	5.30	Summarize ideas	5.75	Summarize ideas	6.29
Edit phrases for interpretation	5.23	Consider syntax or morphology	5.06	Consider syntax or morphology	4.18
Consider syntax or morphology	4.81	Assess task completion or relevance	3.52	Assess task completion or relevance	4.10
Summarize ideas or propositions	3.25	Consider spelling or punctuation	3.04	Assess reasoning, logic, or topic development	2.85
Assess task completion or relevance	2.57	Assess reasoning, logic, or topic development	3.00	Assess text organization	2.81
Assess style, register, or genre	2.21	Assess style, register, or genre	3.00	Assess style, register, or genre	2.73
Assess comprehensibility	2.21	Consider own personal response, expectations, or biases	2.87	Consider spelling or punctuation	2.69

Note. The percentage values have been rounded to the nearest hundredth.

As can be seen from Table 37, overall, all three experience groups used similar strategies when assessing low-quality essays, with the most similarity found between the medium- and high-experienced groups. Medium- and high-experienced raters employed the same top six strategies and shared a total of nine out of ten top rating behaviors. While all three groups shared eight out of ten rating behaviors, the low-experienced group recorded these strategies in an order that differed from the medium- and high-experienced groups, with the strategy, "summarize ideas or propositions" appearing seventh for low-experienced raters but fourth for the more experienced groups. Similarly, while the medium- and high-experienced groups recorded "assess reasoning, logic, or topic development" as a top-ten strategy, the low-experienced group did not use this strategy with the same frequency. Instead, low-experienced raters used the strategies, “edit phrases for interpretation” and “assess comprehensibility,” with greater frequency than their more experienced peers, who did not record these strategies in their top ten. This suggests that low-experienced raters attempted to understand low-quality essays more frequently than more experienced raters, who tended more frequently to assess logic or topic development.

Nonetheless, similarities can be observed in the rating behaviors of each experience group for low-quality essays. All three experience groups relied on the same top three strategies (“read or reread text,” “articulate or revise scoring,” and “read or interpret scoring scale”), and seemed to prioritize a focus on language, with strategies such as, “consider spelling or punctuation” and “consider syntax or morphology” appearing among the top ten decision-making behaviors for low-quality essays. As will be seen in Table 38, these strategies were less commonly used when assessing high-quality papers. Table 38 presents the most commonly used decision-making behaviors by each experience group when assessing high-quality essays.

Table 38

The Most Common Decision-making Behaviors for High-quality Essays

Low-experienced raters		Medium-experienced raters		High-experienced raters	
Decision-Making Behaviors	**%**	**Decision-Making Behaviors**	**%**	**Decision-Making Behaviors**	**%**
Read or reread text	28.40	Read or reread text	25.92	Read or reread text	24.59
Articulate or revise scoring	14.69	Articulate or revise scoring	10.22	Articulate or revise scoring	13.18
Read or interpret scoring scale	14.24	Read or interpret scoring scale	9.81	Read or interpret scoring scale	12.66
Summarize ideas or propositions	5.56	Summarize ideas or propositions	8.05	Summarize ideas or propositions	9.30
Edit phrases for interpretation	4.24	Consider syntax or morphology	4.63	Articulate general impression	5.18
Articulate general impression	3.64	Articulate general impression	4.53	Assess task completion or relevance	3.88
Consider syntax or morphology	3.42	Assess reasoning, logic, or topic development	4.13	Rate ideas or rhetoric	3.55
Assess reasoning, logic, or topic development	2.29	Assess task completion or relevance	3.77	Assess reasoning, logic, or topic development	3.02
Consider spelling or punctuation	2.29	Rate ideas or rhetoric	3.27	Consider syntax or morphology	2.49
Assess coherence	2.14	Assess text organization	2.57	Discern rhetorical structure	2.44

Note. The percentage values have been rounded to the nearest hundredth.

Table 38 compares the top ten most frequently used strategies for each experience group for high-quality essays. When assessing high-quality essays, the three experience groups used the same top four strategies: (1) read or reread text, (2) articulate or revise scoring, (3) read or interpret scoring scale, and (4) summarize ideas or propositions. These four decision-making behaviors accounted for 62.89% of the total strategies used by low-experienced raters, 54.00% of the total strategies used by medium-experienced raters, and 59.73% of the total strategies used by high-experienced raters when assessing high-quality essays. Thus, low-experienced raters tended to rely on these four strategies more than their more experienced peers. This is striking, considering that these four strategies appear as the top four strategies for medium- and high-experienced raters but not for low-experienced raters when assessing low-quality essays.

When comparing the remaining decision-making behaviors used by raters in each experience group, differences can be seen between low-experienced raters and the other two groups. Medium-experienced and high-experienced raters shared nine of their top ten decision-making behaviors. Only the tenth most commonly used strategy for each group was not present in the other's top ten behaviors, with the medium-experienced group using the strategy, "assess text organization," and the high-experienced group using, "discern rhetorical structure," as its tenth most common behavior. The similarity in the most common decision-making behaviors used by medium- and high-experienced raters suggests that these two groups employ similar strategies when assessing high-quality essays. In contrast, the low-experienced raters employed three strategies ("edit phrases for interpretation," "consider spelling or punctuation," and "assess coherence") in their top ten most common behaviors that were not found in the other two groups, suggesting that low-experienced raters rely on different strategies when assessing high-quality essays.

When evaluated collectively, Tables 37 and 38 suggest that medium- and high-experienced raters displayed similar decision-making behaviors, while low-experienced raters differed slightly from these two more experienced groups. Broadly, medium-experienced and high-experienced raters tended to employ the same strategies while rating essays of both low- and high-quality. However, for both low- and high-quality papers, the low-experienced raters seemed to rely on more language-focused strategies, particularly with respect to mechanics. Across experience groups, raters displayed more language-focused strategies—such as considering punctuation, spelling, and syntax—for low-quality essays than high-quality essays.

The data collected from written score explanations provided for each essay were analyzed to triangulate the findings for RQ6. Data were analyzed for each experience group and then compared. Table 39 presents the most commonly identified themes in the three reasons given by low-experienced raters, providing overall frequencies and percentages as well as a breakdown in terms of connotation.

Table 39

The Most Common Reasons Provided by Low-experienced Raters

Theme	**Overall Reasons**		**Positive Reasons**		**Negative Reasons**	
	f	%	*f*	%	*f*	%
Grammar	339	15.75	132	6.13	207	9.61
Organization	308	14.31	165	7.66	143	6.64
Relevance	236	10.96	90	4.18	146	6.78
Mechanics	212	9.85	38	1.76	174	8.08
Topic development	168	7.80	79	3.67	89	4.13
Lexis	140	6.50	89	4.13	51	2.37
Coherence	124	5.76	32	1.49	92	4.27
Register	98	4.55	5	0.23	93	4.32
Comprehensibility	97	4.51	47	2.18	50	2.32
Thesis	90	4.18	37	1.72	53	2.46
Language overall	89	4.13	50	2.32	39	1.81
Content	67	3.11	45	2.09	22	1.02
Introduction	64	2.97	34	1.58	30	1.39
Fluency	48	2.23	19	0.88	29	1.35
Redundancy	39	1.81	1	0.05	38	1.76
Conclusion	24	1.11	4	0.19	20	0.93
Overall impression	9	0.42	6	0.28	3	0.14
Other	1	0.05	0	0.00	1	0.05
Total	**2153**	**100**	**873**	**40.55**	**1280**	**59.45**

Note. The percentage values have been rounded to the nearest hundredth.

The most common reasons provided by low-experienced raters for their essay scores were grammar, organization, relevance, and mechanics. These findings support the findings of the TAPs, in which it was observed that low-experienced raters tended to prioritize language-focused strategies, particularly with respect to mechanics. Mechanics accounted for nearly 10% of the reasons provided by low-experienced raters, and reasons related to mechanics were overwhelmingly negative (8.08%). Furthermore, decision-making behaviors related to grammar and topic development figured prominently in the strategies recorded in the TAPs; however, organization and relevance—two of the top three reasons provided by low-experienced raters—did not emerge as commonly used strategies in the think-aloud protocol data.

Table 40 presents the frequency analysis for the inductively coded themes and connotations for the three reasons provided by the medium-experienced raters.

Table 40

The Most Common Reasons Provided by Medium-experienced Raters

Theme	**Overall Reasons**		**Positive Reasons**		**Negative Reasons**	
	f	%	*f*	%	*f*	%
Grammar	251	15.23	110	6.67	141	8.56
Organization	233	14.14	117	7.10	116	7.04
Topic development	193	11.71	87	5.28	106	6.43
Relevance	161	9.77	74	4.49	87	5.28
Coherence	105	6.37	26	1.58	79	4.79
Content	105	6.37	59	3.58	46	2.79
Language overall	93	5.64	50	3.03	43	2.61
Thesis	81	4.92	44	2.67	37	2.25
Mechanics	78	4.73	9	0.55	69	4.19
Register	75	4.55	12	0.73	63	3.82
Comprehensibility	75	4.55	41	2.49	34	2.06
Lexis	64	3.88	41	2.49	23	1.40
Introduction	53	3.22	32	1.94	21	1.27
Redundancy	20	1.21	0	0.00	20	1.21
Conclusion	19	1.15	5	0.30	14	0.85
Overall impression	19	1.15	15	0.91	4	0.24
Fluency	17	1.03	9	0.55	8	0.49
Other	6	0.36	1	0.06	5	0.30
Total	**1648**	**100**	**732**	**44.42**	**916**	**55.58**

Note. The percentage values have been rounded to the nearest hundredth.

As Table 40 shows, the most commonly given reasons by medium-experienced raters were grammar, organization, topic development, and relevance. These themes largely correspond to the top decision-making behaviors used by medium-experienced raters in the think-aloud protocols, although a greater prioritization of text organization can be seen in the three reasons data (14.14%) than in the TAPs (2.57%).

Continuing the analysis by experience group, Table 41 presents the findings from the three reasons for scoring provided by high-experienced raters.

Table 41

The Most Common Reasons Provided by High-experienced Raters

Theme	**Overall Reasons**		**Positive Reasons**		**Negative Reasons**	
	f	**%**	***f***	**%**	***f***	**%**
Grammar	313	19.27	162	9.98	151	9.30
Organization	263	16.19	189	11.64	74	4.56
Topic development	124	7.64	77	4.74	47	2.89
Language overall	123	7.57	78	4.80	45	2.77
Lexis	108	6.65	67	4.13	41	2.52
Mechanics	107	6.59	30	1.85	77	4.74
Content	103	6.34	74	4.56	29	1.79
Relevance	100	6.16	20	1.23	80	4.93
Thesis	97	5.97	34	2.09	63	3.88
Register	59	3.63	28	1.72	31	1.91
Coherence	57	3.51	23	1.42	34	2.09
Comprehensibility	52	3.20	23	1.42	29	1.79
Fluency	42	2.59	19	1.17	23	1.42
Introduction	33	2.03	18	1.11	15	0.92
Conclusion	20	1.23	3	0.18	17	1.05
Overall impression	13	0.80	7	0.43	6	0.37
Redundancy	7	0.43	0	0.00	7	0.43
Other	3	0.18	2	0.12	1	0.06
Total	**1624**	**100**	**854**	**52.59**	**770**	**47.41**

Note. The percentage values have been rounded to the nearest hundredth.

As can be seen from Table 41, high-experienced raters tended to provide more positive comments (52.59%) than negative comments (47.41%) overall. They were the only experience group to record more positive comments than negative comments, suggesting that high-experienced raters tended to approach essays more positively than their less experienced peers. This data corroborate a major trend in the quantitative data, in which high-experienced raters were found to give higher scores on average than medium- or low-experienced raters.

When themes from the three reasons data are considered individually, a similar trend can be observed. High-experienced raters tended to list grammar and organization as their top reasons for essay scoring. While reasons about grammar were evenly split as positive (9.98%) and negative (9.30%) comments, reasons pertaining to organization were predominantly positive (11.64%; negative = 4.56%). In contrast, medium- and low-experienced raters tended to provide more a balanced distribution of positive and negative comments pertaining to organization and were overall more negative in their reasons related to grammar, again suggesting that high-experienced raters were more likely to prioritize positive aspects, or less likely to emphasize negative aspects, of the essay than their less experienced peers.

After the experience groups were analyzed individually, the findings of the written score explanations analysis were compared across experience groups. Table 42 compares the frequencies with which the 18 thematic categories were found in the three reasons data for each experience group.

Table 42

Comparison of Reasons Provided across Experience Groups

Themes	**Low-experienced raters**		**Medium-experienced raters**		**High-experienced raters**	
	f	**%**	***f***	**%**	***f***	**%**
Coherence	124	5.76	105	6.37	57	3.51
Comprehensibility	97	4.51	75	4.55	52	3.20
Conclusion	24	1.11	19	1.15	20	1.23
Content	67	3.11	105	6.37	103	6.34
Fluency	48	2.23	17	1.03	42	2.59
Grammar	339	15.75	251	15.23	313	19.27
Introduction	64	2.97	53	3.22	33	2.03
Language overall	89	4.13	93	5.64	123	7.57
Lexis	140	6.50	64	3.88	108	6.65
Mechanics	212	9.85	78	4.73	107	6.59
Organization	308	14.31	233	14.14	263	16.19
Overall impression	9	0.42	19	1.15	13	0.80
Redundancy	39	1.81	20	1.21	7	0.43
Register	98	4.55	75	4.55	59	3.63
Relevance	236	10.96	161	9.77	100	6.16
Thesis	90	4.18	81	4.92	97	5.97
Topic development	168	7.80	193	11.71	124	7.64
Other	1	0.05	6	0.36	3	0.18
Total	**2153**	**100**	**1648**	**100**	**1624**	**100**

Note. The percentage values have been rounded to the nearest hundredth.

A comparison of the three reasons provided by raters across experience groups suggests that medium- and high-experienced raters tended to focus on content (6.37%; 6.34%, respectively) more often than low-experienced raters (3.11%). This finding supports the results of the TAPs analysis. Moreover, in line with the findings of the TAPs analysis, low-experienced raters focused more on mechanics (9.85%) than medium- (4.73%) and high-experienced raters (6.59%).

Unlike in the TAPs analysis, organization emerged as a major theme in the three reasons data. The medium- and high-experienced groups also seem to diverge more in the distribution of their reasons provided than in the distribution of their decision-making behaviors recorded during TAPs. While similar themes emerged in the data sets for both groups, their frequency distribution—particularly in comparison to the low-experienced group—is more diverse and less consistent in the three reason data than in the TAPs.

Finally, to summarize the findings from the three reasons analysis, grammar and organization were the most commonly provided reasons for essay scoring, followed by relevance and topic development. Low-experienced raters were found to prioritize mechanics more than medium- and high-experience raters, and high-experienced raters tended to focus on more positive aspects of the essays while scoring than medium- or low-experienced raters did.

Summary of the Findings

This section aims to summarize the findings pertaining to each research question. Table 43 illustrates the findings belonging to RQ1 and RQ2 while Table 44 summarizes the findings for RQ3. Following that, Table 45 presents the summary findings related to RQ4; Table 46 gives a brief summary for the results of RQ5 and RQ6; and Table 43 depicts the results of the first two research questions derived from the descriptive and inferential statistics.

Table 43

Summary Findings for RQ1 and RQ2

Research Question	Result
Are there any significant differences among the analytic scores of the low- and high-quality EFL essays?	Yes. The analytic scores assigned to high-quality and low-quality essays differed from each other significantly, with higher scores assigned to high quality essays.
Are there any significant differences among the analytic scores assigned by raters with varying previous rating experience?	Yes. High- and low-experienced raters' total scores assigned to low-quality essays differed from each other significantly, with high-experienced raters giving higher scores. Additionally, these two groups' ratings assigned to the mechanics component of low-quality essays showed significant differences.

In response to RQ1, the statistical analysis revealed significant differences in the scores assigned to high-quality and low-quality essays. The score differences were observed for each participant rater, with raters assigning significantly higher scores to high-quality essays.

As for RQ2, the descriptive statistical analysis showed that higher scores were assigned to the essays as the rating experience level of the raters increased. The inferential statistics revealed significant differences between low- and high-experienced groups in terms of their total ratings assigned to low-quality essays. Further analysis showed that significant differences between these two groups were observed in the ratings of three essays. Moreover, these two groups' ratings to the mechanics component of the low-quality essays displayed significant differences. With respect to the total scores and components' scores assigned to high-quality essays, no significant differences were found between three experience groups.

In the analysis conducted based on the experience levels that raters self-described rather than reported as their actual experience, the reverse results were obtained in that self-described low-experienced raters differed from self-described high-experienced raters in their total ratings assigned to high-quality essays. However, there were no significant differences regarding the total scores assigned to low-quality essays between any of the self-described experience groups. More strikingly, rater groups based on their self-described experience showed significant differences in their scores assigned to each subscale of the analytic rubric for high-quality essays. The scorings assigned by self-described experience groups to three rubric components were significantly different for low-quality essays.

Table 44 summarizes the findings pertaining to the third research question. These results were obtained from generalizability analysis and separate G-studies were conducted for the three sets of essays—all essays, high-quality essays, and low-quality essays. The aim of this piece of analysis was to estimate the variance components and their relative contributions to the variation of analytic scoring.

Table 44

Summary Findings for RQ3

Research Question	**Result**
What are the sources of score variation that contribute most (relatively) to the score variability of the analytic scores of EFL essays?	In *p x r x q* design, the largest variance component was the object of measurement (45.3%) followed by the residual (31.3%). The interaction of rater and essay quality revealed the third largest variance (14%) and the interaction of persons and rater caused a variance of 7.6%. A small portion of variance occurred due to the raters (1.6%). While the interaction of persons and essay quality had a very small impact on the score variability (0.2%), essay quality did not contribute to the variance at all.
	In *p x r* designs for high- and low-quality essays, the largest variance component was the residual (54.2% and 46.2%, respectively), indicating that most of the variance components cannot be explained. The variability due to raters was larger for high-quality essays (39%) compared to the impact of rater facet in the variability of scores assigned to low-quality essays (23.4%). The high-proficient persons (students) were quite homogeneous (6.8%) in their writing abilities whereas the differences observed between less proficient students were larger (30.4%).

In both G-study designs in Table 44, it was contended that the contribution of the residual to the score variance was large. In other words, the variance components due to the multiple interactions of the facets and the existence of unsystematic and systematic variance sources are not explained. The writing task, on the other hand, distinguished between low-proficient and better students in terms of their writing abilities. However, students who were considered weak showed more differences in their writing proficiency compared to high-proficient student writers. Finally, raters contributed to the variability of scores within the ratings of high-quality and low-quality essays whereas collectively, the variance component due to the rater facet was smaller.

Table 45 illustrates the G- and dependability coefficients based on the essay quality and the rater experience group.

Table 45

Summary Findings for RQ4

Research Question	**Result**
Does the reliability (e.g., dependability coefficients for criterion-referenced score interpretations and generalizability coefficients for norm-referenced scores interpretations) of the analytic scores of raters differ based on their amount of experience?	The generalizability and dependability coefficients were high for the ratings of the essays combining different qualities within rater experience groups. In addition, the analysis revealed high G- and Φ indices for low-quality essays for each rater experience group. However, raters pertaining to low-, medium-, and high-experienced group were considerably more inconsistent for high-quality essays. Especially high-experienced raters displayed greater variability with respect to dependability coefficient compared to less-experienced raters.

As can be understood from Table 45, raters were consistent while grading the essays collectively. Moreover, they did not differ from each other when it came to the evaluation of low-quality essays. However, the ratings assigned to high-quality essays differed markedly, indicating differences between raters within each group in terms rating leniency and severity for high-quality essays.

Table 46 gives the main points of the findings derived from the primary qualitative data (TAPs), which were analyzed using deductive coding.

Table 46

Summary Findings for RQ5 and RQ6

Research Question	Results
How do raters make decisions while rating different quality EFL essays analytically?	-Overall, more interpretation strategies were used than judgement strategies. - Low-quality essays attained slightly more interpretation strategies while raters used slightly more judgement strategies for high-quality essays. -The most commonly used strategies belonged to self-monitoring focus followed by rhetorical and ideational focus and language focus, respectively. -Raters focused more on style, grammar, and mechanics when rating low-quality essays but more on ideas, rhetoric, and their general impression of the essay when rating high-quality essays.
How is rating experience related to EFL raters' decision-making processes and the aspects of writing they attend to?	-Medium- and high-experienced raters displayed similar decision-making behaviors, while low-experienced raters differed slightly from these two more experienced groups. -Low-experienced raters used more interpretation strategies than their more experienced peers whereas medium- and high-experienced raters employed judgment strategies more frequently than raters with less experience did. -Medium-experienced and high-experienced raters tended to employ the same strategies while rating essays of both low- and high-quality essays. -For both low- and high-quality papers, the low-experienced raters seemed to rely on more language-focused strategies, particularly with respect to mechanics. -Across experience groups, raters displayed more language-focused strategies—such as considering punctuation, spelling, and syntax—for low-quality essays than high-quality essays.

The findings presented in this chapter are discussed in more detail as well as with respect to existing literature in the following chapter.

Chapter V

Discussion and Conclusion

This chapter starts with a discussion of the findings obtained from this research study emphasizing the factors—essay quality and rater experience—on the reliability of analytic scoring and decision-making strategies. Following that, the limitations of the study are presented to be considered for future research. Then, in light of the discussion and limitations, implications are discussed from both a pedagogical and methodological perspective followed by a conclusion section and considerations for future research.

Discussion

This research investigated the impact of essay quality and rating experience on the analytic ratings of essays through mixed-method analysis. This dissertation analyzed the scores assigned to essays by raters with varying experience. In addition, the research focused on the decision-making behaviors of the raters while assessing essays of low- and high-qualities. The findings belonging to each research question are related and discussed as follows: The findings pertaining to RQ1 and RQ2 are discussed together since they were connected and they were both analyzed using descriptive and inferential statistics. The second set of questions—RQ3 and RQ4—are combined in the discussion because the findings pertaining to these questions were derived from generalizability analysis. Finally, RQ5 and RQ6 are discussed together since their findings are both based on the qualitative data.

Discussion of findings for RQ1 and RQ2. The essays used in the study were collected from first year students enrolled in an ELT Department at a state university. They were then divided into two distinct qualities—high and low—for the purpose of the study. As for the research participants, raters were full-time EFL instructors at a state university except two participants who were employed as research assistants at the ELT Departments of their universities. Before they started the assessment task in the study, the raters were informed about the student background and their expected proficiency level.

Although the raters did not know the quality division in the essay pack, they were able to give significantly different scores to the essays of two distinct qualities. As suggested by Freedman (1984), using an analytic rubric helped raters distinguish proficient writers from novice writers. However, the score range for almost every essay was very high, which might be the result of the contrast effect in that raters tend to give higher or lower scores to a composition after assessing a better or worse composition in terms of quality (Daly & Dickson-Markman, 1982; Freedman, 1981; Hughes & Keeling, 1984). Moreover, raters knew that the compositions were written by pre-service EFL teachers, which might have increased some of the raters' expectations, resulting in very low scores for low-quality papers or lower scores than deserved for some high-quality essays. Diederich (1974) asserted that raters tend to assign higher scores to the same essays when they are told that the essays are written by better students. In addition, rating context might explain the large score ranges observed for the ratings of both high- and low-quality essays. Baker (2010) found that, although raters were able to distinguish between low- and high-quality essays, they tended to assign different scores to the same essays under authentic (high-stakes) and research (low-stakes) conditions. Thus, the scores may have differed if they had been assigned under authentic conditions.

The raters scored low-quality essays more consistently and they showed greater variation when they rated the high-quality essays. This finding was contradictory to the findings of previous research (Han, 2017; Huang et al., 2014), in which raters were reported to be more consistent while grading papers of higher qualities. These results indicate that text quality has an impact on the reliability of scorings but score variations among raters based on text quality might reveal contrasting findings depending on study context, raters' background, or their expectations in accordance with the students' profile. To make it clearer, raters in this study agreed on the proficiency level of EFL essays with poorer text features and gave lower scores to these essays, which is parallel with the previous studies (e.g. Engber, 1995; Ferris, 1994; Han, 2017; Huang et al., 2014; Russikoff, 1995; Song & Caruso, 1996; Vaughan, 1991); they, however, fluctuated in their understanding of better writing abilities, as more variation was found in the scores assigned to high-quality essays. This variation among raters underlines the fairness problem in EFL assessment for high-proficient student writers in that raters' idiosyncratic constructs and beliefs for a good quality essay might be different even when they are guided by detailed criteria. Another salient reason for this variation might be related to the expectations that raters had set prior to the assessment task. In this regard, some raters might have engaged with the students' performance emotionally and they might have expected more from their prospective colleagues in terms of their writing abilities.

When previous rating experience was considered, it was found that raters with more experience tended to assign higher scores and raters with less experience tended to assign lower scores, suggesting that more experienced raters assessed the essays more leniently compared to their less experienced peers. The qualitative findings corroborate this pattern in that it was found that as the experience level increased, the attitudes that raters developed towards the essays were more positive. However, this does not necessarily imply that relatively more positive qualitative evaluations result in significantly higher scores, as found by Shi's (2001) study in which no significant differences were observed in the scores assigned by raters, despite differences in the qualitative judgements made by the two groups. The percentage of positive comments across rater experience groups were 40.55% for low-experienced raters, 44.42% for medium-experienced raters, and 52.59% for high-experienced raters unlike the findings of previous research (Barkaoui, 2010a), which reported that more experienced raters were more negative compared to novice raters. In addition, previous research showed that raters with varying experience did not differ in their analytic scorings, yet more experienced-raters gave higher scores holistically (Song & Caruso, 1996). Additionally, in contrast to the findings of this research, previous research has suggested that less experienced raters tend to give higher scores (Rinnert & Koyabashi, 2001) and be more lenient in their analytic scorings (Barkaoui, 2011a; Sweedler & Brown, 1985). However, other research has suggested that inexperienced raters tend to be more severe than experienced raters (Weigle, 1999), supporting the findings of this study. The findings that this research arrived at might be explained by the more experienced raters' text repertoire, which they have built up by assessing diverse written performances, resulting in more realistic judgements.

Raters with similar experience are more likely to base their judgements on shared criteria of writing proficiency because they tend to conceptualize L2 proficiency in similar ways (Erdosy, 2004). In this sense, higher levels of inter-rater reliability coefficients were found for all essays mixed with high- and low-qualities together within each experience group including low-experienced raters (.93), medium-experienced raters (.92), and high-experienced raters (.93). Similarly, these three groups showed similarities within themselves while grading low-quality essays (.85, .79, .85, respectively). However, lower coefficients (.44, .46, .27, respectively) were obtained for all experience groups' ratings to high-quality essays. Moreover, the inconsistency that high-experienced group displayed was considerably higher compared to their less-experienced peers. This might have resulted from more-experienced raters' tendency to adhere less to the scoring rubric and greater dependency on their expectations (Eckes, 2008), which may also account for their relatively higher scores overall. Because it is less likely for more experienced raters to shift away from the criteria they have been used to and to adopt new criteria for their assessments (Cumming et al., 2002), they may be more likely to rely on their own expectations. If their expectations differ, such as due to differences in their individual experiences, the scores assigned by high-experienced raters may differ as well, accounting for the greater inconsistency among high-experienced raters. In other words, they might have relied on their self-criteria while assessing essays in higher qualities more than the other experience groups did. However, in contrast to what the current study revealed, Cumming (1990) found that expert raters scored more consistently. Importantly, it should be noted that this research did not compare novice and experienced raters as Cumming (1990) did; rather, raters with varying previous experience were under investigation in this study.

When the analyses on the total scores assigned by the experience groups were examined, it was found that high-experienced raters and low-experienced raters differed from each other significantly in their ratings for low-quality essays. However, rater groups did not record significant differences in terms of their scores to high-quality essays, despite the fact that greater variance was found for high-quality essays. Another important finding was the difference between rater experience groups regarding their scores assigned to analytic rubric components in that low- and high-experienced raters gave significantly different scores to the mechanics component of the scoring scale while assessing low-quality essays. The qualitative data findings explain this difference to some extent: raters with less experience (5.30%) attended to the mechanical aspects of the essays more frequently compared to their more experienced peers (medium-experienced = 3.04%; high-experienced = 2.69%). From this finding, it can be understood that the way raters make use of the given scale and how they prioritize some of the criteria may cause dissimilar ratings, suggesting that unless the raters are trained to use the scoring criteria effectively, using a rubric might not make much of a difference compared to criteria-free assessments (Rezaei & Lovorn, 2010).

Interestingly self-described rater groups (i.e. perceived experience) did not differ significantly in their scores assigned to low-quality essays, unlike the findings obtained from the comparison between reported experience (i.e. experience in years) groups. However, significant differences were found in the scores assigned to high-quality essays across self-described experience groups, with self-described high-experienced raters assigning significantly higher scores than self-described low-experienced raters. Moreover, between self-described experience groups, statistically significant differences were found for all rubric components for high-quality essays and the scores assigned to three rubric components (grammar, style and quality of expression, and mechanics) for low-quality essays. These findings suggest that self-described or perceived experience may influence scoring to a greater extent than actual years of experience.

Moreover, Lim (2011) suggests that practice in the form of rating volume can reduce differences in scoring behaviors between expert and novice raters, and Leckie and Baird (2011) suggest that raters become more homogenous as they rate more essays. Because experience in years may not necessarily reflect the amount of practice that a rater has, measures of self-described experience might be a better indicator of rating experience since raters might have based their perceptions on their previous practices. In other words, some raters might have evaluated a large number of essays over a short period of time while other raters might have assessed a smaller number of essays over a longer period of time, suggesting further investigation is needed on the relationship between actual experience in years and self-described experience based on raters' perceptions.

To summarize, essay quality and rater experience played a role on the ratings of EFL writing. Firstly, all the raters differed in their scores given to low-quality and high-quality essays, meaning that the raters could distinguish between essays of high- and low-quality. Secondly, low-experienced and high-experienced groups showed statistically significant differences in their total scores and the sub-scores that they assigned to the mechanics component of low-quality essays, with high-experienced raters assigning higher scores than low-experienced raters. Thirdly, every rater experience group was inconsistent within itself while grading high-quality essays, suggesting that raters are less consistent in the scores assigned to high-quality than low-quality essays. Finally, self-described experience—in other words, how raters perceive themselves regarding their experience—seemed to have a greater impact on the variability of essay scores, suggesting avenues for future research. To put it differently, experience in writing assessment may not be a sufficient criterion for drawing a reliable rater profile but affective factors such as the raters' motivation and perceptions of themselves as EFL assessors may be essential for rater reliability as well.

Discussion of findings for RQ3 and RQ4. As for the findings derived from generalizability analysis, the largest variance component (45.3%) was due to persons (*p*) when the essay qualities are considered within the same design (*p x r x q*), indicating that students, as intended, differed in their writing performance as measured by the writing task. However, when the analyses were carried out for high-quality and low-quality essays, the persons facet explained a relatively smaller portion of variation especially for high-quality essays (6.8%) compared to the variance observed in low-quality essays (30.4%). Smaller variations due to persons in individual designs were expected given that the designs included more homogenous student groups. However, it was evident that students were more dissimilar from each other in the lower-proficiency group in terms of EFL writing abilities, suggesting a larger performance band among the lower-quality essays. This might have resulted from the essay quality division carried out by the expert raters in that they might have treated some of the medium-quality essays as though they belonged to the lower proficiency level.

In *p x r x q* G-study design, the second greatest variance was attributable to the residual (31.3%), which was obtained from the interaction of raters, compositions, essay quality, and other unexplained unsystematic and systematic sources of errors. This amount was considerably higher in the scorings of high-quality and low-quality essays (54.2% and 46.2%, respectively). These findings signify the existence of other factors contributing to the score variation such as writing task, scoring method, raters' educational background, expectations, gender, etc., in the measurement design (Brennan, 2001a; Huang, 2008; Huang et. al., 2014).

Due to the limited number of facets in the designs, high effects on the score variability were expected from the residual.

With regards to the contribution of rater facet (*r*) to the variation of analytic scores in three designs, raters varied substantially more in terms of leniency and severity while grading high-quality essays (39%) compared to their contributions to the scorings of low-quality essays (23.4%). However, when quality was considered collectively, the raters' impact on the score variability was much smaller (1.6%). These findings suggest that raters are more consistent when assessing a pile of essays of mixed or differing writing proficiency levels than when assessing essays of a similar quality level (e.g., only high-quality or only low-quality essays). In other words, consistent scoring can be expected in large-scale assessment contexts in which students from varying proficiency levels exist. However, rater consistency regarding leniency and severity seems to be a concern when it comes to interpreting essays of a certain quality, especially for written performances with better text features. That is to say, when presented with a set of high-quality essays, raters become less consistent in their scoring. Thus, EFL instructors may benefit from a rater-training program for the conceptualization of quality constructs in EFL compositions.

As for the other variance components for the collective scorings of high- and low-quality essays, 14% of the total variance was due to the interaction between raters and essay quality, indicating that some raters differed substantially while scoring compositions of distinct qualities. Another variance source was attributable to the interaction of persons and raters (7.6% of the total variance), referring to the inconsistency between certain raters in terms of severity and leniency while assessing certain essays. The remaining variance sources including essay quality (0%), and person-by-quality (0.2%) did not seem to influence the variability of scores.

Despite large variances that stemmed from the residual, high dependability coefficients were obtained from the scorings of mixed-quality and low-quality essays across the three rater groups. This might be related to the size of the rater facet in that a large number of facets leads to higher dependability indices (Brennan, 2001a; Güler et. al., 2012; Shevelson & Webb, 1991). Thus, the number of raters were decreased to observe the most practical assessment situation, and it was found that three raters for mixed-quality essays and 10 raters for low-quality essays would produce acceptable dependability coefficients (i.e. above .80). However, as for the ratings of the high-quality essays, 73 raters should be needed to arrive at dependable scores, indicating a larger contribution of the rater facet to the variability of the scores. Decreasing the number of raters for ratings of separate designs for low-quality and high-quality essays would result in lower reliability and generalizability coefficients, which raise a potential concern about the fairness of ratings (Huang, 2008). Stated more plainly, while three raters would be sufficient to ensure reliable scores for a set of mixed-quality essays, 10 raters would be necessary to ensure reliable scores for a set of low-quality essays and 73 raters would be necessarily for a set of high-quality essays. These results suggest that in order to ensure reliability in high-stakes EFL writing assessment contexts at tertiary education, such as entrance and exit tests in the English preparatory programs or writing exams for the selection of students for international exchange programs, double-grading or even engaging three raters in the evaluation process seems to be necessary for fair scorings. However, assessing students' writings pertaining to the same proficiency level (e.g. proficiency levels in English preparatory programs such as beginner, intermediate, advanced, etc.) may be problematic in that realistic scenarios could not be reached for the ratings of certain quality essays isolated from diverse proficiency levels in terms of writing abilities, particularly when the writing samples are expected to all be high-quality composition. However, although administering only one or two tasks to the students at the same time because of the practical considerations is suggested

(Weigle, 1999), so as to maximize the reliability of ratings in writing performance assessments, increasing the number of tasks would be more cost-efficient than increasing the number of ratings per task (Baker, 2012; Lee, Kantor, & Mollaun, 2002). In this regard, testing students writing ability with multiple tasks or topics in Turkish tertiary educational contexts may be an effective way of providing fairer judgements to the students, particularly to groups of similar- rather than mixed-ability students.

In conclusion, the writing task partitioned the students in terms of their writing abilities when high- and low-proficient students were considered collectively. When high-quality essays were analyzed separately, the fluctuation among students were considerably smaller compared to the low-quality essay writers, indicating a bigger range among lower proficient students in their writing performance. On the other hand, larger rater variation was observed considering the ratings for high-quality essays compared to those of low-quality essays. Larger contributions of the residual in all designs indicated the impact of hidden facets in the variability of scores regardless of the essay quality. Finally, the variance due to the interaction of raters and essay quality indicated that some raters substantially differed from their peers while assigning their scores to essays of distinct qualities and varied in their scorings while assessing certain essays as evident from the contribution of the rater-persons interaction as a variance component. Based on the aforementioned discussion, this study highlights important considerations to improve EFL writing assessment practices for low-stakes and high-stakes tests at institutional and large-scale contexts.

Discussion of findings for RQ5 and RQ6. The qualitative data included think-aloud protocols and the written score explanations that raters gave to justify their total scores assigned to each essay. Collectively, the coded strategies obtained from the protocols showed that the raters uttered slightly more interpretation strategies (53.98%) than judgement strategies (46.02%), similar to Cumming et. al.’s (2002) and Gebril and Plakans’ (2014) findings but unlike Barkaoui's (2010b) findings. A similar pattern was observed when the strategies were examined separately for each essay quality. However, low-quality essays attained slightly more interpretation strategies while raters used slightly more judgement strategies for high-quality essays, indicating that raters endeavored more to comprehend what low-quality essays wanted to say. These differences show that the proficiency of a composition affects the decision-making strategies of the scorers (Cumming et al., 2002). Considering the raters by their experience level, low-experienced raters attended to interpretation strategies more frequently compared to their more experienced peers, who reported relatively more judgement strategies.

When the strategies were examined by focus, the most commonly used strategies belonged to self-monitoring focus (59.33%) followed by rhetorical and ideational focus (23.95%) and language focus (16.72%) respectively, which corroborates the findings of previous research (Barkaoui, 2010b; Gebril & Plakans, 2014). The same trend was evident both for high-quality and low-quality essays. Nonetheless, raters used slightly more strategies in self-monitoring/rhetorical and ideational foci for high-quality essays while low-quality essays attracted more language-related strategies compared to high-quality essays. Cumming et al. (2002) and Gebril and Plakans (2014) found a similar tendency in their studies, in which raters attended to the linguistic features of low-proficient essays more when compared to better essays. This shows that raters prioritized form more than content in low-quality essays. However, Han (2017) concluded a reverse pattern in his study in that papers regardless of their quality attracted more language focus strategies followed by rhetorical/ideational focus and self-monitoring focus, respectively. Additionally, in this study, raters' previous experience seemed to affect their strategy preferences as well. While low-experienced raters used more self-monitoring focus and language-related strategies, more experienced raters attended to the strategies related to the rhetorical and ideational focus more than low-experienced raters did, which overlaps with the findings of Barkaoui's (2010b) study.

The most commonly used strategy for all essays was “read or reread text” (27.98% of the total strategies), followed by “articulate and revise scoring” (11.91%) and “read or interpret scoring scale” (11.41%), all of which belong to the self-monitoring focus. These strategies were also defined as the most frequently used decision-making behaviors in Barakaoui’s (2010b) study. Given that all raters were expected to read an essay at least once, refer to the rubric for their evaluations, and assign a score to the essay in the end, these results are not surprising. In addition, using an analytic rubric might make these strategies more dominant since raters considered different aspects of the essay in accordance with the rubric components, leading to use the aforementioned strategies multiple times for each essay. These three behaviors accounted for 56.27% of all strategies used by the low-experienced group, 45.98% of the strategies used by the medium-experienced group, and 50.06% of all strategy use by the high-experienced group. As suggested by Wolfe et al. (1998), these numbers indicate that low-experienced raters might have displayed a more bottom-up approach to the assessment task while more experienced raters may have evaluated the essays in a more holistic manner. Adopting different approaches for interpreting the scoring scale might lead to use of different decision-making behaviors (Barkaoui, 2010b). Nevertheless, across the three experience groups, seven out of ten of the top ten strategies used were the same. In terms of commonalities, medium- and high-experienced raters differed only in that medium-experienced raters recorded, “consider personal response, expectations, or bias,” as the tenth most common strategy, while “rate ideas or rhetoric” appeared in the top ten strategies for high-experienced raters. The fourth most commonly strategy overall was “summarize ideas or propositions” (6.10%), indicating that raters interacted with the essay content to improve their comprehension.

When considering the remaining six strategies found in the top ten for each essay quality, “consider syntax and morphology” and “consider spelling and punctuation,” were more commonly used for low-quality essays compared to high-quality essays, which is also supported by Gebril and Plakans (2014), who stated that language becomes a more predominant and decisive feature for raters to assign their scores while evaluating low-proficient texts. This finding triangulates the quantitative findings in that the only statistically significant difference regarding component scores was found between low- and high-experienced raters' scores assigned to the mechanics component of low-quality essays. The strategy, "assess style, register, or genre," was also more commonly used for low-quality essays, listed ninth in the frequency list, and it did not appear in the top ten most common strategies for high-quality essays; rather, it ranked 19th for high-quality essays. This might be related to the less proficient students' tendency towards benefiting from their L1 in generating and interpreting texts in L2. In addition to the English language proficiency of the students, other factors such as their L1, home culture, and style of written communication can affect writing performance (Hinkel, 2003; Yang, 2001) and impact rater behaviors exhibited during ESL writing assessments (Bachman, 2000). It should be noted that text problems related to direct translation from Turkish were considered within scale of "style and quality of expression" in the scoring criteria used in this research, perhaps contributing to more references to style in the TAPs for low-quality essays. When raters attended to translation issues in the essays during the verbal protocols, the researcher assessed these utterances within the behavior of "assess style, register, or genre."

The strategies "articulate general impression" and "rate ideas or rhetoric" were more commonly used for high-quality essays than low-quality essays, suggesting that less language-related concern about the text enables raters to shift their focus to the development of ideas in the text, as supported by the findings of Gebril and Plakans (2014). Considered collectively, these trends suggest that raters focused more on style, grammar, and mechanics when rating low-quality essays but more on ideas, rhetoric, and their general impression of the essay when rating high-quality essays. In other words, while assessing low-quality essays, raters interacted with three rubric components—grammar, style and quality of expression, and mechanics—more frequently, and they focused more on the content and organizational aspects of the rubric while grading high-quality essays.

When the written score explanations given for the essays were examined, the overwhelming majority of the reasons for high-quality essays was positive (70.66%), while 79.88% of reasons for low-quality essays were negative, similar to the findings of Barkaoui's (2010a) study. These percentages support the findings obtained from the scores, in which high-quality essays were given higher scores which were reflected in the raters' attitudes towards the essays. As for high-quality essays, organization and grammar were the two most commonly given reasons, followed by topic development, relevance, and lexis. These findings corroborate the data collected from the think-aloud protocols in that syntax and morphology, task completion and relevance, and topic development were found to be common strategies used to assess high-quality essays. However, data derived from the written explanations found a much strong emphasize on organization and lexis than emerged from the decision-making behaviors used by raters to grade high-quality essays. This might have resulted because raters were expected to prioritize three reasons to justify their scores over others compared to the verbal protocols in which raters had the flexibility of saying aloud what came to their minds related to the essays, resulting in slight differences in data sets obtained from TAPs and written score explanations.

When the written score explanations for low-quality essays were analyzed, grammar, organization, and mechanics were found to be the most common reasons. While 10.67% of reasons for low-quality essays related to mechanics, this category accounted for less than 4% of reasons for high-quality essays. The difference in attention to mechanics for high- and low-quality essays supports the findings of the think-aloud protocols, in which it was found that raters tended to focus more on elements of language such as grammar, spelling, and punctuation when assessing low-quality essays than high-quality essays. Similarly, while 5.99% and 10.72% of reasons related to content and topic development respectively for high-quality essays, these two themes accounted for only 4.16% and for 7.17%, respectively, of the reasons provided for low-quality essays, corroborating the findings from the think-aloud protocol data that suggest that raters tend to focus more on ideas when assessing high-quality essays.

All raters regardless of experience frequently considered the aspects of grammar and organization. However, a comparison of the three reasons provided by raters across experience groups suggests that medium- and high-experienced raters tended to focus on content (6.37%; 6.34%, respectively) more often than low-experienced raters (3.11%). Moreover, low-experienced raters focused more on mechanics (9.85%) than medium- (4.73%) and high-experienced raters (6.59%). Similarly, Barkaoui (2010a, 2010b) found that novice raters attended to the mechanics component more than their experienced peers did. These results indicate that raters' thinking processes and the aspects they attend to more frequently might lead to scoring differences as observed in this research for the ratings of mechanics component of the rubric. In other words, raters may arrive at different analytic scores based on somewhat different qualitative criteria (Shi, 2001). In order to minimize the inconsistency between rater groups, orienting them to consider as many aspects as possible in accordance with the scoring guidelines and essay features might be an effective model of rater training.

As a result, raters' decision-making strategies and their attitudes towards students' writing abilities were related to the essay quality and their previous rating experience. Low-quality essays attracted slightly more interpretation strategies compared to high-quality essays for which raters used slightly more judgement strategies. Furthermore, for low-quality essays raters focused on language-related aspects more frequently that they did during the scoring of high-quality essays. It was evident from their written explanations that more experienced raters were more positive towards the students' written performances, which was reflected in their ratings. Moreover, low-experienced raters paid more attention to mechanics in the essays, leading them to differentiate from the more-experienced raters; this was reflected in the significantly different scores assigned to the mechanics component. Overall, the discussion presented in this section reveals differences between raters, which might be helpful to understand the variability of EFL writing scores and to design a rater training and even a more detailed scoring criteria to reduce the fluctuation between the ratings.

Limitations of the Study

The score variations between and within rater groups might be related to the lack of rater training. Although the scoring rubric was developed with the participant raters, they were not given a thorough rater training but were simply oriented to the scoring scale. As such, training raters together (group training) on how to use the rubric and discussing rating standards to assess essays of different qualities might have resulted more reliable scores and decreased the range of scores assigned by raters (Attali, 2015; Barkaoui, 2010b).

Moreover, the researcher did not control the rating time and conditions in this research; rather participants were allowed to carry out the assessment task at their homes to minimize the pressure of the researcher. However, differences in rating times and conditions (e.g. fatigue, personal issues, raters' emotional status etc.) might have affected inter-rater reliability (Barkaoui, 2010b). Future research could attempt to control for these differences.

In this study, raters scored the essays written on a single topic and their scoring performance might have been affected by the writing topic or prompt. Previous research has shown that writing task and essay topic can impact the variability of essay scores (Gebril, 2009; Hamp-Lyons & Mathias, 1994; Jennings et al., 1999; Saeidi & Rashvand Semiyari, 2011; Weigle, 1999). In this sense, although the researcher allowed the students to develop background knowledge about the topic and gave sufficient time to write the essays in order to minimize the impact of writing task, topic, and prompt, students might have performed differently with a different writing topic or prompt. In addition, raters' scoring performance might have been different with another topic.

In the same vein, G-studies revealed large residual variance for mixed quality (31.3%), high-quality (54.2%), and low-quality (46.2%) essays, indicating that other facets hidden in the residual might have contributed to the large score variance (Brennan, 2001a; Huang et. al., 2014). This is not a desirable situation in generalizability analysis and it is assumed that smaller contribution of the residual as a variance component increases the possibility of explaining explicit variance sources in G-studies.

As part of qualitative data, raters were expected to list their explanations to justify their scores given to the essays. Previous research relied on this type of data in holistic assessment contexts (e.g. Barkaoui, 2010a; Milanovic, Saville, & Shuhong, 1996; Rinnert & Koyabashi, 2001). Using an analytic scoring scale in this research might have hidden the self-criteria of the raters for their justifications; rather, raters might have relied on the analytic rubric descriptors to give their written explanations for the assigned scores.

Think-aloud protocols as a data collection tool are another limitation in that five raters in this research failed to provide TAPs to the researcher due to several reasons. In spite of the detailed guidelines and the sample-training video, 15% of the verbal protocols were not recorded in the required format and content. Further, although high degree of agreement was achieved between coders (.83 with $p < .001$), coding raters' utterances might be problematic since they might have meant something else than what the researcher interpreted (Cumming et al., 2001, 2002). Additionally, the demonstration of a sample think-aloud on a writing assessment task might have biased raters in their decision-making behaviors (Cumming et al., 2002). Additionally, thinking aloud might have put pressure on raters, resulting in problems with the quality and quantity of verbalization (Barkaoui, 2011a, 2011b). Furthermore, as suggested by Barkaoui (2010a, 2010b, 2011a), the experimental nature of the study might have affected the scores and the thinking processes of the raters as the assessment task lacked real-life conditions related to educational and assessment contexts.

Conclusion

First, the essays used in this study were comprised of two distinct qualities and raters were not informed about the quality division. The statistical analysis revealed that raters showed statistical differences in their scores assigned to high-quality and low-quality essays. In other words, all raters were able to distinguish low-proficient student authors from their high-proficient peers.

Second, the statistical analysis showed that raters varied from each other in their ratings based on their previous rating experience. High-experienced raters and low-experienced raters displayed statistically significant differences in their total ratings of low-quality essays. Furthermore, statistically significant differences were observed between their sub-scores assigned to the mechanics component of the low-quality essays. When the scoring pattern across experience groups was examined, a positive correlation between the average scores and the amount of rater experience was seen in that more experienced raters gave higher scores to the essays than low-experienced raters did.

Third, G-theory analysis revealed that the variance component due to rater facet was considerably high when the ratings of high-quality and low-quality essays were evaluated separately. However, the score variability due to raters was much smaller collectively, indicating that raters showed great differences in terms of leniency and severity within each essay quality than in the overall mixed-quality set. In addition, G-theory analyses were not able to explain a considerable amount of variance in all G-study designs (e.g. *p x r x q* for all papers, *p x r* for high- and low-quality papers) due to the residual variance component. The residual contains multiple interactions and other systematic and unsystematic sources. For example, this study did not consider the topic or rating method, which might have affected the score variability.

Fourth, an almost perfect degree of inter-rater reliability was achieved within each rater group for low-quality and mixed-quality (high- and low- quality papers together) essays and D-studies showed that a lower number of raters would still produce scores with an acceptable level of dependability index. However, the reverse is true for high-quality essays in that low dependability coefficients were found across the three rater groups, and only if the number of raters were increased unreasonably would reliable scores be obtained for high-quality essays.

Finally, raters displayed different decision-making strategies based on essay quality and rating experience. More experienced raters were more positive compared to less experienced raters, leading to higher essay scores respectively. Generally, raters used more interpretation strategies than judgement strategies. However, language-related strategies were more frequently used for low-quality essays than high-quality essays. The high-quality essays attracted strategies under the self-monitoring/rhetorical and ideational foci more frequently compared to low-quality essays. While medium-experienced and high-experienced raters were similar in their decision-making behaviors, low-experienced raters had their own pattern in terms of rating behaviors in addition to moderate commonalities with the other two rater groups.

Pedagogical Implications

In light of the findings and limitations of this study, several pedagogical implications can be drawn. Primarily, this study did not compare novice and expert raters but investigated the differences between raters with varying rating experience. The variations between raters in their scorings show that experienced raters might have their own standards and revise them in time as they become more experienced. Instead of presuming experienced raters to be more reliable markers, the findings underline the need for a detailed and continuous rater training even for raters with extensive rating experience. In this way, scoring gaps can be reduced between raters to arrive at fair judgements.

Another implication addresses in-house rating protocols and rater training for the teachers working in the same institution. Fifteen of the participants of this study were from the English Preparatory Program of Bursa Technical University. The program has a double-grading system for writing performance assessments throughout the academic year (“Quality Manual”, 2015); nevertheless, no protocol is available for matching the pairs for grading the performances. Considering the findings of this research, this program and all other English programs can consider matching relatively high-experienced and relatively low-experienced raters in pairs for double-grading students’ writing, as high-experienced raters were found to be more lenient compared to their less-experienced peers. In other words, if two experienced raters are paired, they are more likely to give higher scores to a certain writing performance, while the same essay might receive a considerably lower score if the grading is conducted by two less experienced raters. Matching relatively high- and low-experienced raters together could compensate for these effects in double-grading situations.

Moreover, traditional rater training models can be revisited as the findings suggested that score variations between raters may be related to differentiation in certain sub-scores of writing (e.g. mechanics component), given that certain raters (e.g. low-experienced raters) prioritized strategies related to such components (e.g. consider spelling and punctuation) in their think-aloud protocols and the written explanations for their ratings (Cumming et al., 2002). As such, developing a rater-training model that shifts raters’ focus to all aspects of writing covered by the scoring criteria (Eckes, 2008) instead of emphasizing certain traits such as grammar, content, organization, etc., might help ensure intra- and inter-rater reliability and achieve more dependable scores. That is to say, a strategy-based rater training model built upon the most commonly used decision-making strategies may lead raters to think similarly while evaluating EFL compositions, thus resulting in more consistent scores.

Although using an analytic rubric is considered reliable and advantageous over holistic scoring (Charney, 1984; Cohen & Manion, 1994; Cumming, 1990; Elbow, 1999; Hamp-Lyons, 1990; Reid, 1993; Shi, 2001; Weigle, 2002; White, 1994), the findings showed that score variations could be observed during analytic evaluation. As such, rather than using traditional holistic and/or analytic scoring scales, developing a clear and user-friendly scale (Huang & Foote, 2010) with more detailed descriptors (Knoch, 2009) might be helpful to reduce the inconsistencies between the raters. Added to these suggested rubric traits, context-bound scoring scales can be developed considering the local, cultural and institutional dynamics.

Furthermore, evaluating students' performance and raters' scoring performance on a single topic might be limited for the generalizability of the scores since writing topic or prompt is an effective factor in the variability of essay scores (Gebril, 2009; Hamp-Lyons & Mathias, 1994; Jennings et al., 1999; Saeidi & Rashvand Semiyari, 2011; Weigle, 1999). As such, judging students' writing abilities as well as rater' scoring performances across different topics and occasions might minimize these pedagogical concerns.

Finally, this research showed that self-described experience seems to have greater impact on scoring compared to actual rating experience as measured in years. The rater identities that teachers have built might be related to other factors such as theoretical information in EFL writing assessment, teaching experience, or personality factors in addition to previous rating experience. As such, these issues can also be considered as selection criteria for establishing a rater team for in-house or larger assessment contexts. Further research is needed to examine the effects that rater self-perception has on scoring behavior.

Methodological Implications

Moving from the findings and limitations of this study, several methodological implications can be identified. Firstly, investigating teachers' rating performances using only an analytic scale is limited in drawing generalizable conclusions. Instead, comparing the rating performances of the raters in terms of score variability and rater severity, and self-consistency across different types of scoring scales might provide deeper insight into writing assessment research (e.g. Bacha, 2001; Barkaoui, 2007b, 2010c; Han, 2013; Knoch, 2009; Weigle, 2002). Additionally, examining raters' decision-making strategies across different types of assessment criteria (e.g. holistic, analytic etc.) in the same research context might provide a better understanding of how raters' decision-making behaviors evolve with different assessment criteria.

Another implication related to the methodological aspect is about the conditions under which student writers generate their essays and raters evaluate these texts. In both cases, the researcher wanted to provide flexible environments to avoid pressure on both students and raters, which could stem from the existence of the researcher and the research context. However, simulating a real-life context or relying on a naturalistic context might reveal different results. Students might approach the writing task more seriously under exam-like conditions, and controlling raters to avoid differences in rating time and conditions (e.g. fatigue, personal issues, raters' emotional status etc.) might produce better results in terms of dependability (Barkaoui, 2010b).

The quantitative framework that guided this study was the G-theory approach to determine the sources of score variability and dependability of the scores. It was helpful to underline the impact of raters in the reliability of scores but not sufficient to determine the hidden facets contributing to the score variations. As such, including multiple facets (e.g. essay topic, prompt, rating methods, etc.) can produce a better picture of the multiple variance components and their relative contributions to the score variability. In other words, investigating the reliability of ratings on multiple topics (e.g. two essay type, topics or prompts) through multiple rating method (e.g. holistic and analytic) by multiple rater groups (e.g. novice and expert raters) might be a better design to determine as many variance sources and their relative contributions to score variability as possible.

Another implication can be drawn about the use of verbal protocols in writing assessment research. In this study, the researcher used a sample training video on how to conduct TAPs and prepared detailed guidelines in order to train raters to verbalize their thoughts. However, it was evident that theoretical training was not sufficient to ensure a complete understanding or adequate execution from all raters. Instead, practical or hands-on training may be necessary to increase the quality and quantity of data from think-aloud protocols (e.g. Han, 2017). For example, raters might be asked to assess one sample essay while thinking-aloud and they can discuss their experience with the researcher to become better oriented to verbalize their thoughts.

Future Research

Based on its findings, limitations and implications, this study underscores four research areas for future studies. Firstly, analytic scoring is expected to produce reliable scores because it limits the raters to consider the aspects of writing covered by the scoring scale and controls the score weights allocated for each trait (Goulden, 1994). However, raters regardless of their previous marking experience showed great variation in their analytic ratings while assessing essays written by better students in terms of writing abilities. As such, future research can consider examining the impact of raters' assessment standards and their expectations on the analytic and holistic scoring of compositions written by high-proficient L2 authors since raters seemed to use their own criteria rather than adhering to the scoring scale in this research.

Secondly, the analytic scoring scale used in this study was adapted with the participant raters, considering that they would be more likely to base their judgements on the scoring scale if they were involved in the scale-development and adaptation processes (Barkaoui, 2007b; Davidson, 1991; Hamp-Lyons, 1991; Weigle, 2002). However, this procedure limited the raters only to consider the given traits and their score weights. In the end, the dependability of the scores especially for high-quality essays were considerably low for each rater group, raising concerns about raters' holistic approach even when using an analytic rubric. Although they are less practical, analytic scoring scales are preferred over holistic scales to ensure reliability (Goulden, 1984; Perkins, 1983), yet this preference is likely to bring validity problems to the assessment task since raters have more flexibility to consider different aspects of writing in holistic assessment rather than limiting themselves to the given aspects an analytic scoring rubric. As such, future research can consider involving raters in the development of a scoring scale that combines analytic and holistic approach traits from the very beginning instead of adapting an existing one.

Thirdly, as the findings suggested, how raters perceive themselves in terms of experience is an important consideration in addition to actual experience. The greater impact of raters' self-described experience on the variability of essay scores can be further investigated using both qualitative and quantitative research. Uncovering the role of and the reasons behind teachers' self-perceptions as raters can shed light on rater reliability issue in EFL/ESL writing assessment.

Finally, it should be noted that fairness is a big problem in EFL writing assessment in Turkey even if certain assessment protocols are adapted for developing and evaluating performance tests. To illustrate, students are subject to learn English in intensive English programs before they start their education in their EMI departments in Turkish tertiary education. Although central authorities regulate higher education policies in Turkey, decisions are made about students' academic careers by individual institutions based on their own assessment practices. With this in mind, this study provides avenues for future research to investigate the reliability of essay ratings for institutional and national large-scale assessments, with specific considerations for individual rater training and writing assessment practices to improve reliability. In doing so, instead of implementing traditional rater training programs that focus on the basics of appropriate use of scoring criteria for rating the essays, the impact of strategy-based rater training, which might integrate commonly used decision-making behaviors and detailed scoring criteria to address the goals and needs of particular writing performance assessment contexts, can be examined.

References

Attali, Y. (2015). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115.

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*(1), 125-141.

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing, 18*, 191-208. doi:10.1016/j.jslw.2009.05.003

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*(3), 371-383.

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

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Bachman, L. F., & Palmer, A. (2010). *Language testing in practice*. Oxford: Oxford University Press.

Baker, A. B. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing, 15*(3), 133-153.

Baker, A. B. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly, 9*(3), 225-248.

Barkaoui, K. (2007a). Participants, texts, and processes in second language writing assessment: A narrative review of the literature. *The Canadian Modern Language Review, 64*(1), 97-132.

Barkaoui, K. (2007b). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86-107.

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes.* Unpublished doctoral dissertation, University of Toronto, Canada.

Barkaoui, K. (2010a). Do ESL essays raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31-57.

Barkaoui, K. (2010b). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing, 27*(4), 515-535.

Barkaoui, K. (2010c). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54-74.

Barkaoui, K. (2011a). Effects of marking method and rater experience on ESL scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279-293.

Barkaoui, K. (2011b). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51-75.

Barrett, S. (2001). The impact of training on rater variability. *International Educational Journal, 2*(1), 49-58.

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching practice: Evaluating assessment essays. *College Composition and Communication, 37*(3), 315-327.

Baydin, A. G. (2006). *Blank map of Republic of Turkey's provinces* [Digital image]. Retrieved from https://commons.wikimedia.org/wiki/File:BlankMapTurkeyProvinces.png

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. London: SAGE.

Breland, H. M. (1983). *The direct assessment of writing skill: A measurement review (ETS Research Report No: 86-9)*. Princeton, NJ: Educational Testing Service.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339-353.

Brennan, R. L. (2001a). *Generalizability theory.* New York, NY: Springer.

Brennan, R. L. (2001b). *Generalizability theory: Statistics for social science and public policy*. New York: Springer-Verlag. Retrieved from https://www.google.com.tr/search?hl=tr&tbo=p&tbm=bks&q=isbn:0387952829

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1-21.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of Psychology, 52*(1), 13-15.

Brown, H. D. (2004). *Language assessment: Principles and classroom practice*. New York, NY: Pearson/Longman.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25*(4), 587-603.

Brown, J. D. (1996). *Testing in Language Programs.* Upper Saddle River, NJ: Prentice Hall.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill College.

BTU SFL (2014). *Holistic scoring scale*. Bursa Technical University School of Foreign Languages, Turkey.

Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and nonnative speakers of English* (*ETS*

Research Report Series GRE Board Research Report GREB No. 83-2R). Princeton, NJ: Educational Testing Service.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing. *Research in the Teaching of English, 18*(1), 65-81.

Cohen, L., & Manion, L. (1994). *Research methods in education.* New York, NY: Routledge.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 compositions really mean? *TESOL Quarterly, 29*(4), 762-765.

Cooper, P. L. (1984). The assessment of writing ability: A review of research *(ETS Research Report Series GRE Board Research Report GREB No. 82-15R)*. Princeton, NJ: Educational Testing Service.

Cresswell, J. W. (2011). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). New Delhi, India: PHI Learning Private.

Cronbach, L. J., Gleser, G. C., Nada, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: Wiley.

Cumming, A. (1989). Writing expertise and second language proficiency. *Language Learning 39*(1), 81-141.

Cumming, A. (1990). Expertise in evaluating second language composition. *Language Testing, 7*(1), 31-51.

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype tasks: An investigation into raters' decision making and development of a preliminary analytic framework (TOEFL Monograph Series, Report No: 22).* Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67-96.

Daly, J. A., & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement, 19*(4), 309-316.

Davidson, F. (1991). Statistical support for training in ESL composition rating. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 155-164). Norwood, NJ: Ablex.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*(1), 7-29.

Diederich, P. B. (1974). *Measuring growth in English*. Urbana, IL: National Council of Teachers of English.

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

Ebel, R., & Frisbie, D. A. (1991). *Essentials of educational measurement.* Englewood Cliff, NJ: Prentice Hall.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270-292.

Elbow, P. (1999). Ranking, evaluating, and liking: Sorting out three forms of judgements. In R. Straub (Ed.), *A sourcebook for responding to students writing* (pp. 175-196). Cresskill, NJ: Hampton Press.

Elorbany, R., & Huang, J. (2012). Examining the impact of rater educational background on ESL writing assessment: A generalizability theory approach. *Language and Communication Quarterly, 1*(1), 2-24.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study for experienced raters of ESL compositions* (TOEFL Research Report RR-03-17). Princeton, NJ: Educational Testing Service.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*(2), 414-420.

Frederiksen, J. R. (1992, April). *Learning to "see:" Scoring video portfolios or "beyond the hunter-gatherer" in performance assessment*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Freedman, S. W. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English, 15*(3), 245-255.

Freedman, S. W. (1984). The register of student and professional expository writing: Influences on teachers' responses. In R. Beach & S. Bridwell (Eds.), *New directions in composition research* (pp. 334-347). New York, NY: Guilford Press.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75-98). New York, NY: Longman.

Fulcher, G. (2010). *Practical language testing*. London: Routledge.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London and New York, NY: Routledge.

Gao, X., & Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education, 14*(2), 191-203.

Gambetti, E., Fabbri, M., Bensi, L., & Tonetti, L. (2008). A contribution to the Italian validation of the General Decision-making Style Inventory. *Personality and Individual Differences, 44*(4), 842-852.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*(4), 507-531.

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing, 15*(2), 100-117.

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing, 21*(2), 56-73.

Goulden, N. R. (1992). Theory and vocabulary for communication assessments. *Communication Education, 41*(3), 258-269.

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *The Journal of Research and Development in Education, 27*(2), 73-82.

Greenberg, K. L. (1992). Validity and reliability issues in the direct assessment of writing. *WPA: Writing Program Administration, 16*(1-2), 7-22.

Güler, N., Uyanık, G. K., & Teker, G. T. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi Yayınları.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). New York, NY: Cambridge University Press.

Hamp-Lyons, L. (1991). Basic concepts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 5-15). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1995). Rating nonnative rating: The trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762.

Hamp-Lyons, L. (1996). The challenges of second language writing assessment. In E. White, W. Lutz & S. Kamusikiri (Eds.), *Assessment of writing: Policies, politics, practice,* (pp. 226-240). New York, NY: Modern Language Association.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing, 3*(1), 49-68.

Hamp-Lyons, L., & Zhang, B. W. (2001). World Englishes issues in and from academic writing assessment. In L. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes*, (pp. 101-116). Cambridge: Cambridge University Press.

Han, T. (2013). *The impact of rating methods and rater training on the variability and reliability of EFL students' classroom-based writing assessments in Turkish universities: An investigation of problems and solutions.* Unpublished doctoral dissertation, Atatürk University, Turkey.

Han, T. (2017). Scores assigned by inexpert raters to different quality of EFL compositions, and the raters' decision-making behaviors. *International Journal of Progressive Education, 13*(1), 136-152.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 1-28). Mahwah, NJ: Lawrence Erlbaum Associates.

Henning, G. (1991). Issues in evaluating and maintaining an ESL writing assessment program. In L. Hapm-Lyons (Ed.), *Assessing second language writing in academic contexts*, (pp. 279-292). Norwood, NJ: Ablex.

Hinkel, E. (1994). Native and nonnative speakers' pragmatic interpretations of English texts. *TESOL Quarterly, 28*(2), 353-376.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly, 37*(2), 275-371.

Homburg, T. J. (1984). Holistic evaluation of ESL composition: Can it be validated objectively? *TESOL Quarterly, 18*(1), 87-108.

Huang, J. (2007). *Examining the fairness of rating ESL students' writing on large-scale assessments.* Unpublished doctoral dissertation, Queen's University, Canada.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? —A generalizability theory approach. *Assessing Writing*, *13*(3), 201-218.

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies*, *5*(1), 1-17.

Huang, J. (2011). Generalizability theory as evidence of concerns about fairness in large-scale ESL writing assessments. *TESOL Journal, 2*(4), 423-443.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17*(3), 123-139.

Huang, J., & Foote, C. J. (2010). Grading between lines: What really impacts professors' holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly, 7*(3)*,* 219 – 233.

Huang, J., & Han, T. (2013). Holistic or analytic – A dilemma for professors to score EFL essays? *Leadership and Policy Quarterly, 2*(1), 1-18.

Huang, J., Han, T., Tavano, H., & Hairston, L. (2014). Using generalizability theory to examine the impact of essay quality on rating variability and reliability of ESOL writing. J. Huang & T. Han (Eds.), *Empirical* quantitative research in social sciences*: Examining significant differences and relationships*, (pp. 127-149). New York, NY: Untested Ideas Research Center.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hughes, A., & Lascaratou, C. (1982). Competing criteria for error gravity. *ELT Journal, 36*(3), 175-182.

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

Huot, B. A. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*(2), 201-213.

Huot, B. A. (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* (pp. 206-236). Cresskill, NJ: Hampton Press.

Hyland, K. (2003). *Second language writing*. New York, NY: Cambridge University Press.

James, C. (1977). Judgements of error gravities. *ELT Journal, 31*(2), 116-124.

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: An investigation of the effect of topic on language-test performance. *Language Testing, 16*(4), 426-456

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485-505.

Kane, M. (2008, November). *Errors of measurement, theory, and public policy.* Paper presented at the 12th Annual William H. Angoff Memorial Lecture. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/PICANG12.pdf

Kenyon, D. (1992, February). *Introductory remarks at symposium on development and use of rating scales in language testing*. Paper presented at the 14th Language Testing Research Colloquium, Vancouver, British Columbia.

Kieffer, K. M. (1998, April). *Why generalizability theory is essential and classical test theory is often inadequate?* Paper presented at the Annual Meeting of the South Western Psychological Association, New Orleans, LA.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.

Knoch, U., Read, J., & Randow, J. V. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26-43.

Kobayashi, H., & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language Learning, 46*(3), 397-437.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly, 26*(1), 81-112.

Krapels, A. R. (1990). An overview of second language writing process research. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (37-56). New York, NY: Cambridge University Press.

Kroll, B. (1990). *Second language writing: Research insights for the classroom*. New York, NY: Cambridge University Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics 33*(1), 159-174.

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418.

Lee, Y.-W., Kantor, R., & Mollaun, P. (2002, April). *Score dependability of the writing and speaking sections of new TOEFL.* Paper presented at the annual meeting of National

Council on Measurement in Education, New Orleans, LA. Abstract retrieved from ERIC. (*ERIC No. ED464962)*

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543-560.

Linn, R. L., & Burton, E. (1994). Performance-based assessments: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13*(1), 5-8.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3)*,* 246-276.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. New York, NY: Peter Lang.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Mackey, A., & Gass, S. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

McColly, W. (1970). What does educational research say about the judging of writing ability? *The Journal of Educational Research, 64*(4), 148-156.

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

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

Milanovic, M., Seville, N., & Shuhong, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic & N. Seville (Eds.), *Performance testing, cognition, and assessment: Selected papers from the 15th Language Testing Colloquium (LTRC), Cambridge and Arnhem* (pp. 92-114). Cambridge: Cambridge University Press.

Myers, M. (1980). *A procedure for writing assessment and holistic scoring*. National Council of Teachers of English, Urbana, IL.

Najimy, N. C. (1981). *Measure for measure: A guidebook for evaluating students' expository writing.* National Council of Teachers of English, Urbana, IL.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly, 17*(4), 651-671.

Plakans, L. (2007). *Second language writing and reading-to-write assessment tasks: A process study*. Unpublished doctoral dissertation. The University of Iowa.

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

Quality Manual (2015, 15 September). *Bursa Technical University School of Foreign Languages Quality Manual.* Retrieved from http://depo.btu.edu.tr/dosyalar/ydyo/Dosyalar/SFL%20-%20Quality%20Manual%20-%2023%2011%202017%284%29.pdf

Raimes, A. (1990). The TOEFL test of written English: Causes for concern. *TESOL Quarterly, 24*(3), 427-442.

Reid, J., & O'Brien, M. (1981, March). *The application of holistic grading in an ESL writing program.* Paper presented at the Annual Convention of Teachers of English to Speakers of Other Languages. Detroit, MI. (ERIC Document Reproduction Service No. ED 221 044).

Reid, J. M. (1993). *Teaching ESL writing*. Englewood Cliffs, NJ: Prentice-Hall.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39.

Rinnert, C., & Kobayashi, H. (2001). Differing perceptions of EFL writing among readers in Japan. *The Modern Language Journal, 85*(2), 189-209.

Russikoff, K. A. (1995, March). *A comparison of writing criteria: Any differences?* Paper presented at the Annual Meeting of the Teachers of English to Speakers of Other languages, Long Beach, CA.

Saeidi, M., & Rashvand Semiyari, S. (2011). The impact of rating methods and task types on EFL learners' writing scores. *Journal of English Studies, 1*(4), 59-68.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate ESL compositions. A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129-152). Cambridge: Cambridge University Press.

Sakyi, A. A. (2003). *A study of the holistic scoring behaviors of experienced and novice ESL instructors.* Unpublished doctoral dissertation, University of Toronto, Canada.

Santos, T. (1988). Professors’ reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22*(1), 69-90.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*(1), 1-30.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing, 14*(2), 157-184.

Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and Psychological Measurement, 55*(5), 818-831.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of educational Measurement, 30*(3), 215-232.

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

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325.

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.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking, and ESL students? *Journal of Second Language Writing, 5*(2), 163-182.

Spicer, D. P., & Sadler-Smith, E. (2005). An examination of the general decision making style questionnaire in two UK samples. *Journal of Managerial Psychology, 20*(2), 137-149.

Stalnaker, J. M., & Stalnaker, R. C. (1934). Reliable reading of essay tests. *The School Review, 42*(8), 599-605.

Suen, H. (1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum.

Sweedler-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluation. *English Journal, 74*(5), 49-55.

Şahan, Ö. (2016a, June 20). *Rubric orientation session* [Video file]. Retrieved from https://www.youtube.com/watch?v=AKPnsdt4nuo

Şahan, Ö. (2016b, June 23). *A sample think-aloud protocol* [Video file]. Retrieved from https://www.youtube.com/watch?v=hoJxNZFdT4Q

Şahan. Ö., & Razı, S. (2017, June). *The impact of rater experiences and essay quality on rater behavior and rating scores*. Paper presented at the 16th Symposium on Second Language Writing, Assessing Second Language Writing, Bangkok, Thailand.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3-12.

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. 11-126). Norwood, NJ: Ablex.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2)*,* 263-87.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2)*,* 145-178.

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

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*(1), 27-55.

Weir, C. J. (2005). *Language testing and validation.* Hampshire: Palgrave McMillan.

White, E. M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance*. San Francisco, CA: Jossey-Bass Publishers.

Wolfe, E. F., & Feltovich, B. (1994, April). *Learning to rate essays: A study of scorer cognition.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Wolfe, E. W. (2005). Uncovering rater's cognitive processing and focus using think-aloud *protocols. Journal of Writing Assessment, 2*(1), 37-56.

Wolfe, E. W., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication*, *15*(4), 465-492.

Yang, Y. (2001). *Chinese interference in English writing: Cultural and linguistic differences.* (ERIC Document Reproduction Service No. ED 461 992).

Appendices

Appendix A: Rater Profile Form

The purpose of this questionnaire is to gather background information for this study. Please note that the aim of the research is not to judge you, but rather to better understand and interpret your writing assessment performance. In addition to other data provided for the study, your information and identity will be kept confidential.

I would like to thank you for your cooperation and contribution to this study.

Your Name (Pseudonym) is……………….

1. Your gender:

Male………. Female……….

2. Your age:

20-30……… 31-40…….. 41-50…….. More than 50…………

3. Your academic rank:

Research Assistant…………. Instructor………….. Assist. Prof. Dr.…………

Other, please specify…………..

4. What is your highest level of education?

B.A……….. M.A………… Ph.D.……….. Other, please specify………..

5. How long have you been teaching EFL?

None..… 2 years or less…… 3 to 4…….. 5 to 6…….7 to 10……..More than 10 years……..

6. How long have you been teaching EFL at the university level?

None..….2 years or less…… 3 to 4…….. 5 to 6…….7 to 10……..More than 10 years……..

7. How long have you been teaching EFL writing?

None..…. 2 years or less …… 3 to 4…….. 5 to 6…….7 to 10……..More than 10 years……..

8. How long have you been teaching EFL writing at the university level?

None..…. 2 years or less …… 3 to 4…….. 5 to 6…….7 to 10……..More than 10 years……..

9. What is your experience marking EFL papers?

I have ….. years experience.

None..…. 2 years or less …… 3 to 4…….. 5 to 6…….7 to 10…….. More than 10 years…….

10. Have you ever received any training in writing assessment?

Yes…… No……..

11. How would you describe your experience as an EFL writing assessor?

I have no experience……..

I have little experience……..

I have some experience……..

I am experienced……..

I am very experienced ……..

Appendix B: Analytic Scoring Rubric

Rater’s Name: Your Score: ...... / 10.0

Essay Code:

Score and Criteria					
	0-0.2	0.3-0.5	0.6-0.9	1.0-1.2	1.3-1.5
Grammar	Many recurring errors in syntax and morphology, resulting in ungrammatical sentences that hinder meaning. Numerous errors such as tense, subject-verb agreement, article, and preposition errors, run-on sentences and fragments, sentence structure errors. The text is not clear or understandable.	Errors in syntax and morphology that hinder meaning. Frequent errors such as incorrect use of auxiliaries and modals, tense agreement, article, and preposition errors, fragments and run-on sentences, difficulty forming complex sentences. The text is difficult to understand and meaning is often lost.	Moderate-level accuracy in syntax and morphology. Some errors such as occasional incorrect use of auxiliaries and modals, tense agreement, articles and prepositions, fragments and run-on sentences. Some local errors in forming complex sentences. The text is mostly clear with some meaning loss.	Generally accurate use of syntax and morphology. Few errors such as incorrect uses of auxiliaries and modals, tense agreement, article and preposition errors. Rare use of fragments or run-on sentences. A few local errors in forming complex sentences. The text is generally clear and understandable.	Almost no errors with syntax and morphology. Almost no errors such as incorrect use of auxiliaries, modals, tense agreement, article or preposition errors. Correct use of complex sentences. The text is clear and understandable.
	0-0.5	0.6-1.1	1.2-1.8	1.9-2.4	2.5-3.0
Content	Off-topic. The text contains almost no elaboration of a single topic or introduces multiple topics. Almost no evidence of a thesis statement and supporting details. Lacks evidence of critical thought.	Wanders off-topic. Ideas are loosely connected and underdeveloped. Major problems with thesis statement and supporting details. Poor evidence of critical thought.	Generally on-topic. Ideas are not fully developed. Some problems with thesis statement and supporting details. Occasionally includes irrelevant details. Some evidence of critical thought.	On-topic. Ideas are generally developed. Almost no problems with thesis statement and supporting details. A few irrelevant details are given. Satisfactory evidence of critical thought.	On-topic. Ideas are fully developed. Clear thesis statement and relevant supporting details. Strong evidence of critical thought.
	0-0.4	0.5-0.9	1.0-1.5	1.6-2.0	2.1-2.5
Organization	Almost no introduction or conclusion. The body of the text lacks unity and cohesion. Almost no organization.	Poor introduction and conclusion. The body of the text lacks organization and transition between ideas. Weak unity and cohesion. Poor logical organization of paragraphs.	Fair introduction and conclusion. The body of the text partly lacks flow of ideas, appropriate transitions, and clear supporting ideas. Some issues with unity and cohesion. Some issues with logical organization of paragraphs.	Clear introduction and conclusion. The body of the text mostly includes clear supporting ideas and transitions. Almost no issues with unity and cohesion. Almost no issues with logical organization of paragraphs.	Exemplary introduction and conclusion. The body of the text includes clear supporting ideas and transitions. Demonstrates exemplary unity and cohesion. Logical organization of paragraphs.

Continues…

	0-0.3	0.4-0.7	0.8-1.2	1.3-1.6	1.7-2.0
Style and quality of expression	Many language errors that interfere with meaning. Many direct translations from Turkish. Weak and inappropriate vocabulary. Unrelated and repetitive sentences.	Frequent language errors and direct translations from Turkish. Limited and repetitive vocabulary. Lacks sentence variety.	Some language errors and direct translations from Turkish. Moderate use of vocabulary. A few repetitive sentences.	Few language errors and direct translations from Turkish. Appropriate and varied vocabulary. Sufficient sentence variety.	Exemplary language use with almost no errors. A wide range of advanced vocabulary. Exemplary sentence variety.
	0-0.1	0.2-0.3	0.4-0.6	0.7-0.8	0.9-1.0
Mechanics	Numerous and recurring spelling, punctuation, and capitalization errors that interfere with meaning.	Many spelling, punctuation, and capitalization errors that occasionally interfere with meaning.	Some spelling, punctuation, and capitalization errors that rarely interfere with meaning.	A few spelling, punctuation, and capitalization errors that do not interfere with meaning.	Almost no spelling, punctuation, or capitalization errors.

(adapted from Han, 2013)

Appendix C: Assessment Instructions for Quality Check Raters

Dear Rater,

I am currently working on my Ph.D. thesis at Çanakkale Onsekiz Mart University in the English Language Teaching Department. The purpose of this research study is to investigate the impact of rater experience and essay quality on rater behaviors and scoring. In this respect, you are kindly requested to assess the essays and categorize them based on their quality, which will serve as a guiding division for the main data collection of the study. Please pay attention to the following items while evaluating the papers in the folder. I would like to thank you in advance for your valuable contribution. With my best regards.

Özgür ŞAHAN
Assistant Director
Bursa Technical University
School of Foreign Languages
e-mail: ozgursahan66@hotmail.com

Essay Topic: Some people think that English teachers working at primary schools and high schools are insufficient to teach English effectively. Therefore, Ministry of Education in Turkey is thinking of hiring native English-speaking teachers to support English language education. Do you think that English teachers in Turkey are qualified enough for teaching English to the students or should English language education in Turkey be supported by native English-speaking teachers? Use specific reasons and examples to develop your essay.

- Please read the essay topic written above before assessing the papers.
- The essays were collected from first-year students enrolled in an ELT Department at a state university.
- Please be aware that the students' English language proficiency level is B1/B2.

- Students were expected to write 500- to 700-word essays.
- The essays were submitted using *Turnitin* to avoid plagiarism incidents.
- You are NOT expected to score the essay. Use the holistic rubric to determine the quality of the paper. There are three quality categories: LOW, MEDIUM, and HIGH. Please indicate the quality by checking one of the boxes at the top of each essay.
- Please evaluate the essays individually rather than comparing them to other essays in the set.
- Please use the rubric for each essay to help standardize the division among the quality categories.

Appendix D: Holistic Scoring Rubric

SCORE & LEVEL	TOPIC DEVELOPMENT	SCORE & LEVEL	LANGUAGE USE
3 **PROFICIENT to ADVANCED** A writing sample that demonstrates **competence** in Topic Development	• Treatment of the topic is relevant, logical and well-developed. • Explanations, examples and/or details related to the topic are satisfactory. • Essay is generally well organized and well-formatted. • Structure of the essay is cohesive. • Content and scope are accurate.	**3** **PROFICIENT to ADVANCED** A writing sample that demonstrates **competence** in Language Use	• Language of the essay flows smoothly. • Use of a variety of structures and expressions; although minor errors/mistakes may exist. • Considerable use of content-based and appropriate vocabulary. • Conventions of the written language are generally correct.
2.5		**2.5**	
2 **PARTIALLY PROFICIENT** A writing sample that suggests **lack of competence** in Topic Development	• Treatment of the topic is somewhat irrelevant. • Explanations, examples and/or details do not fully cover the trend. • Essay is inadequately organized and formatted. • Content and scope are limited and/or inaccurate. • There exists somewhat divergence from the topic.	**2** **PARTIALLY PROFICIENT** A writing sample that suggests **lack of competence** in Language Use	• Language of the essay is generally comprehensible, but it occasionally needs some interpretation on the part of the reader. • Frequent grammatical errors may exist; there may be some redeeming features, such as correct advanced structures. • Limited vocabulary. • Frequent interference from another language may appear in the use of idioms, expressions, statements, etc. in the target language. • Frequent errors in conventions of the written language may be present.
1.5		**1.5**	
1 **UNSATISFACTORY - POOR** A writing sample that demonstrates **lack of competence** in Topic Development	• Essay is a restatement or paraphrasing of the topic. • There is very limited written output. • No explanation, example and/or detail is given. • Essay is poorly-organized. • Content and scope are very limited and/or inaccurate. • There exists excessive divergence from the topic.	**1** **UNSATISFACTORY - POOR** A writing sample that demonstrates **lack of competence** in Language Use	• Insufficient, irrelevant, and very elementary vocabulary. • Prevalent errors in conventions of the written language may interfere with the written communication. • Poor control of grammar debilitates and/or blocks written communication. • Incomprehensible and purposeless target language production.
0.5	• Very limited written output comprising a few words. • No organization. • Totally irrelevant content and scope.	**0.5**	• Total failure in the use of language. • Total lack of written communication.
0	• No written output.	**0**	• No language use.
SCORE of the TOPIC DEVELOPMENT SECTION: _____ out of 3	+	SCORE of the LANGUAGE USE SECTION: _____ out of 3 =	TOTAL SCORE of the TWO SECTIONS: _____ out of 6 / _____ out of 100
a) Score 0 out of 6-point ***0 out of 100*** ***b) Score 0.5 out of 6-point*** ***1 out of 100***	***d) Score 1.5 out of 6-point*** ***18 out of 100*** ***e) Score 2 out of 6-point*** ***27 out of 100***	***g) Score 3 out of 6-point*** ***45 out of 100*** ***h) Score 3.5 out of 6-point*** ***54 out of 100***	***a)Score 4.5 out of 6-point*** ***72 out of 100*** ***b)Score 5 out of 6-point*** ***81out of 100***

(developed by BTU SFL, 2014)

Appendix E: Rubric Orientation Instructions for Raters

Dear Rater,

I am currently working on my Ph.D. thesis at Çanakkale Onsekiz Mart University in the English Language Teaching Department. The purpose of this research study is to investigate the impact of rater experience and essay quality on rater behaviors and scoring. Before collecting the main data of the study, I would like to orient you to the 10-point analytic rubric that will be used in this research. In this respect, you are kindly requested to assess three essays using the rubric provided in this pack. Please pay attention to the following items while evaluating the papers. I would like to thank you in advance for your valuable contribution. With my best regards.

Özgür ŞAHAN
Assistant Director
Bursa Technical University
School of Foreign Languages
e-mail: ozgursahan66@hotmail.com

Essay Topic: Some people think that English teachers working at primary schools and high schools are insufficient to teach English effectively. Therefore, Ministry of Education in Turkey is thinking of hiring native English-speaking teachers to support English language education. Do you think that English teachers in Turkey are qualified enough for teaching English to the students or should English language education in Turkey be supported by native English-speaking teachers? Use specific reasons and examples to develop your essay.

- Please read the essay topic written above before assessing the papers.
- The essays were collected from first-year students enrolled in an ELT Department at a state university.
- Please be aware that the students' English language proficiency level is B1/B2.

- The essays were submitted using *Turnitin* to avoid plagiarism incidents.
- Please evaluate the essays individually rather than comparing them to other essays in the set.
- Please use the rubric for each essay to help assign reliable scores.
- Do not negotiate your decision with anyone else.
- Write your scores for each component (e.g. grammar, content, etc.) on the far right column and the total score at the top of the page (your score). *Please see the analytic scoring rubric sample scores form.*
- After assigning your score, please indicate three reasons to justify your decision at the top of each essay.
- When you finish grading the essays, please provide feedback about the rubric by answering the questions on the rubric feedback form.

Appendix F: Rubric Feedback Form

Rater's Name:

1. How would you assess practicality of the rubric? Please put X in the relevant box.

Rubric Component	Bad	Fair	Good	Excellent
Grammar				
Content				
Organization				
Style and quality of expression				
Mechanics				
Overall				

2. There were 5 performance levels for each component reflected in numbers (e.g. 0-0.4, 0.5-0.9, 1.0-1.2, 1.3-1.6, and 1.7-2.0) Were the expressions for each performance level distinctive to identify the essays' strengths and weaknesses?

□ Yes □ No

3. Were the descriptors clear enough to help you make your decision about the essay?

□ Yes □ No

If not, please indicate the specific descriptor below and offer your suggestions to replace or rephrase it (e.g. "off-topic is not clear for the content component, I would say irrelevant topic"):

……………………………………………………………………………………………

……………………………………………………………………………………………

4. The weights of the components are distributed evenly in the rubric (e.g. max. 2 pts. for each component). Do you think the weight of any component should be increased or decreased (e.g. "max. pts. for organization should be 3 and should be 1 for grammar")?

Yes □ No □

If so, please indicate your suggestions below by specifying the component and the max. pts. that should be assigned to it:

…………………………………………………………………………………………………

…………………………………………………………………………………………………

5. If you have any other comments about the rubric, please indicate below:

…………………………………………………………………………………………………

…………………………………………………………………………………………………

Appendix G: Scoring Rubric Orientation and Adaptation Session Report

On June 15th at 11.45 a.m., the researcher organized an orientation session guided by the grades that the participants of this research project gave to three essays and the feedback they provided for the 10-point analytic scoring scale. The purpose of this piloting phase was to familiarize the raters with the scale and obtain their opinions regarding the practicality of the scale by reflecting on the weaknesses and strengths of it. Given physical and scheduling problems, only 15 participants working at Bursa Technical University were able to join the session. However, all of the feedback about the scale obtained from the 33 raters who are participating in this research was discussed in detail. The discussion lasted around 1 hour and was video-recorded to make sure the raters who could not attend the session were well-informed about the discussions and decisions made during the session.

This report aims to summarize the decisions that were made regarding the scoring scale and provides reasons to justify the necessary changes in three categories as follows:

Wording of descriptors:

- It was agreed that ‘excellent’ should be changed to ‘exemplary’ in the highest band of the rubric, as this seemed more appropriate and objective in describing the relevant performance level.
- The raters agreed that ‘no logical organization’ should be changed to ‘poor logical organization’ in the second column of the organization component, as ‘no’ did not match the other descriptors in this column and did not differentiate from the lowest performance band of the organization component.
- It was suggested that ‘good’ in the highest performance bands of the organization and style and quality of expression components should be changed to ‘more than expected;’ however, the suggestion was rejected as the

raters agreed that it was more subjective. Instead, the raters agreed that the adjective ‘exemplary’ should be used for these descriptors to match the first decision made in this meeting.

- In the feedback received on the rubric, some raters suggested that descriptor such as ‘some,’ ‘clear,’ and ‘moderate’ were unclear. This was discussed among the raters in the meeting. Because more objective alternatives could not be proposed, the raters decided to keep the original descriptors.
- One rater suggested that ‘minor issues’ was an unclear descriptor; this was discussed, but the other raters decided that minor issues was clear and any alternatives would appear more subjective to the rater.
- In their feedback, some raters suggested that quantifiers such as numerous, many, some, etc., should be specified with numbers, as this would clarify the descriptors for the raters. However, one rater noted that focusing on the quantity may distract the raters from focusing on the quality. Because the aim of the rating is not error analysis, it was decided that numbers would not be used in the rubric.
- Some of the raters suggested that the use of ‘no’ as a descriptor in the lowest performance level band was ambiguous. One rater noted that ‘no’ deserves 0.0 but the band ranges up to 0.4, presenting problems for grading on the scale. Other raters noted the importance of keeping ‘no’ on the scale. A compromise was reached: ‘no’ will be replaced with ‘almost no’ in the lowest bands of the content and organization components.

Items and Components

- One rater suggested that the organization component should be renamed as ‘organization and cohesion,’ claiming that cohesion is distinct from organization. This was debated among the raters, who did not agree that the category should be renamed. The raters decided that the component should not be renamed but the descriptors within component could be made more clear by expanding 'some organization problems' to include cohesion and flow of ideas. Also, the raters agreed that 'minor issues' and 'almost no issues' in the third and fourth performance levels of the organization component were not distinctive enough and it was decided that these descriptors should be rephrased.

- Many raters suggested that the content component is not comprehensive enough to guide the rater. This was discussed and it was decided that the descriptors of the content component should be strengthen by adding items such as evidence, example, critical thought, thesis statement, supporting details, etc.
- It was suggested that 'spelling' and 'capitalization' should be divided in the rubric; however, the suggestion was not accepted by the raters and no change will be made.
- Similarly, it was suggested that the introduction, body, and conclusion should be handled separately in the rubric. As with the previous suggestion, this suggestion was not accepted by the raters. No change will be made. Further, the raters noted that when the content component is strengthened, this will become clearer.
- Some raters noted that the highest band of the style and quality of expression component should include a descriptor related to 'translations from L1.' However, the other raters agreed that the descriptor of 'exemplary language

use' in the relevant performance level implied that the essay would have authentic language use without translations from L1. No change will be made.

Weight Distribution

- The raters decided that the middle range should be broadened and the two lowest bands should be narrowed. The new distribution among the performance levels for two-point component will be 0-0.3, 0.4-0.7, 0.8-1.2, 1.3-1.6, 1.7-2.0. The distribution among the performance levels for different-weighed components will be balanced accordingly.
- Many raters suggested that the content and organization components should be prioritized in the rubric. The raters decided that the weight distribution of the components should be revised as follows:
    - ✓ Content: 3 pts.
    - ✓ Organization: 2.5 pts.
    - ✓ Style and quality of expression: 2 pts.
    - ✓ Grammar: 1.5 pts.
    - ✓ Mechanics: 1 pt.
- Some raters suggested that a 100-point scale would be more user-friendly than a 10-point scale, as they are more familiar with a 100-point scale. However, the researcher noted that the literature suggests that 10-point scales are more reliable than 100-point scales in terms of rater variation. It was decided that 10-point system will be used.

If you have any questions or comments concerning the decisions, please contact the researcher.

June 18 2016
Özgür ŞAHAN

Appendix H: Instructions for Think-Aloud Protocols

Please read these instructions carefully before you begin assessing the essays.

Purpose

These instructions are written to help guide you and others in producing think-aloud protocols for this project in a consistent and informative manner. Think-aloud protocols ask people to say everything they are thinking while they perform a task in order to document and better understand what raters pay attention to and consider important when they complete a task. The purpose of the think-aloud protocols for this study is to find out in as much detail as possible what you as an assessor of EFL essays are thinking, deciding, and doing while rating a sample of EFL essays. The most important thing to remember is to say everything you are thinking, and to make certain this is recorded clearly onto the voice-recorder. What you say will become important data for my research. Thank you in advance!

The Assessment Task

You will receive a package of 50 written essays produced by first-year university EFL students with B1/B2 language proficiency level. However, you will assess 16 of them using think-aloud protocol while the other 34 essays will be assessed following standard procedures. The essays that you will assess by thinking aloud have been tagged at the top of the page with the notification 'USE THINK-ALOUD PROTOCOL DURING YOUR ASSESSMENT!' Those essays will also be listed at the end of this document. Say as much as you can while you are reading the essay and deciding on how to rate it, and be sure the score you assign to each essay is recorded along with your ongoing impressions of the essay.

The Essays

You will also receive copies of the essay prompt that was originally given to the students so that you will know what they were asked to write about. There is only one essay topic and one essay type.

The essays have been identified with code numbers. The order in which you receive the essays has been sequenced randomly in quality, but you should receive compositions of distinct qualities.

The Ratings

In making your assessment, try to use the analytic rubric as the basis for your decision. The rating will not be judged as right or wrong. However, I will be analyzing the scores that you assign to the essays along with the spoken data regarding your thoughts while assessing the essays.

Recording Your Thoughts While Assessing

- Keep talking, conveying your thoughts continuously while you assess the essays beginning from the moment you first see the essay until you have completed rating it.
- Feel free to speak in either English or Turkish. If you speak in Turkish, it will be translated into English for the final analysis of the data.
- Speak continuously. Report fully, even what might seem trivial. Do not assume that others know what you are doing or thinking.
- Try to avoid speech fillers (i.e., uh, um) as much as possible. Try to use words instead, so that I can understand what your thoughts are.
- Talk and make your assessment as naturally and as honestly as you can, according to what you usually do when you assess students' essays. Don't start rationalizing your ideas at length; I am just interested in your natural thought process as you make decisions.

Instructions for Recording

1. Turn on the voice-recorder or your smart phone so that you can record your voice and check that it works. Check whether it records properly and that the quality of the

recording is clear by trying out a few words initially, then playing it back. Make sure there is no background noise (e.g., fans, music, foot tapping, etc.).

2. Keep the recorder/smart phone at an appropriate distance from your face and be sure it captures your voice clearly.

3. Turn on the recorder/smart phone, and state the essay code and your name (or pseudonym to be used in the research) at the beginning of each essay.

4. While rating the essay, follow the instructions above (*Recording Your Thoughts While Assessing).* Then, when you have made a rating decision, indicate the score (out of 10) that you have assigned to the essay.

5. Report your first impression of the essay and whether it influences your rating. Then continue talking—saying what you think—as you are making your assessment decisions.

6. You will write three reasons that impact you most for your decision about the essay. Feel free to write other notes on the essays if you like, but I will not be analyzing your written notes.

7. You may read the essays aloud or silently according to what feels most “natural” to you. Make sure you report exactly what you are doing. If you are reading silently, indicate which part of the essay elicits your comments.

8. If you happen to reconsider any of your ratings (e.g., for a second or third time), verbalize your reason(s) for doing so and indicate on the recorder that this is what you are doing.

9. If you have to take a break while you are assessing the essays, indicate on the recorder that you are doing this, turn the voice recorder off or pause the recording on your smart phone. Then, when you start again, indicate this clearly on the device.

10. When you have completed assessing the essay, turn off the voice-recorder/smart phone.

11. Please record your thoughts for each essay separately.

12. At the end of the assessment session and voice-recording, please put all the essays together with the recorder that you used (if you were provided with one) back into the package. Thank you!

(adapted from Cumming, Kantor, & Powers, 2001, pp. 83-85)

THINK-ALOUD PROTOCOLS WILL BE USED FOR THE FOLLOWING ESSAYS:

ANK1617
ELZ1625
ERZ1666
ESK1617
GTW1617
KYS1666
MNS1617
NZL1617
NZL1666
NJ1666
SNP1666
SVS1666
TKT1666
TRZ1666
YYS1624
YZT1625

Appendix I: Examples of Coded Decision-Making Behaviors

Self-Monitoring Focus—Interpretation Strategies

1. Read or interpret essay prompt: a) "Let me check the essay prompt again; maybe I remember wrong or this student is confused." (Yakup); b) "What was the essay prompt? Let me check. Well, are Turkish teachers sufficient or should we hire native English speaking teachers? OK." (Tugce)
2. Read or reread text: a) "I am reading this again silently." (Onur); b) "I am reading the essay now." (Arif)
3. Envision personal situation of the writer: a) "This text is so bad. Is this student going to be an English teacher?" (Oznur); b) "What I understand from this essay is that the student is not at a B2 proficiency level." (Hamit)
4. Scan or skim text: a) "Just looking over the text and I see that, uh, he continued writing on the back page." (Mesut); b) "Let me go back to text very quickly to see whether I underlined any problems related to the vocabulary." (Onur)
5. Read or interpret scoring scale: a) "As for the style scale, uh, it has some language errors but not too many so I am going to give 1 point for the style and quality of expression." (Remzi); b) "OK, let's check the rubric and I think the grammar fits the highest score range because there is almost no mistake." (Sertap)

Self-Monitoring Focus—Judgment Strategies

1. Decide on a macrostrategy for reading and rating: a) "I would like to get an idea about the composition by looking at the title first." (Celal); b) "I will read the essay first and then will try to evaluate it with the scoring scale." (Mehmet)
2. Consider own personal response, expectations or biases: a) "I think, he is criticizing the teachers who graduated from the departments other than English language teaching harshly. I graduated from the department of American literature and culture and I kind of

feel offended actually." (Celal); b) "I expect to see a detailed road map in the introduction to guide me for the remaining parts of the essay. I hope I am not being harsh on essays since I am looking for this particular feature." (Adalya)

3. Define or revise own criteria: a) "I think my initial impression is changing as I continue to read the text." (Ahmet); b) "Okay, the essay is over and, in the beginning, I talked about the, the organization and transition between ideas but at the end there, the essay changed my mind." (Bilal)
4. Compare with other compositions or "anchors": a) "This seems to be a longer one compared to the other essays that I have read so far." (Kaan); b) "I believe this is the best thesis statement I have ever seen so far." (Derya)
5. Summarize, distinguish, or tally judgments collectively: a) "It is generally on topic. There is no problem with grammar. I think, the organization is not bad as well." (Nazende); b) "First of all, the organization is not good. The essay starts with an example that should be given in the body part. Also, the essay did not touch upon the topic at all. There are not many problems with the expressions, yet I can see some serious grammar problems and a few spelling mistakes." (Kayra)
6. Articulate general impression: a) "The essay is not very strong, actually." (Ayten); "Well, this is a good one." (Guney)
7. Articulate or revise scoring: a) "I am giving 2 points for the content." (Ayten); b) " Overall, this essay gets a 9 out of 10." (Celal)

Rhetorical and Ideational Focus—Interpretation Strategies

1. Interpret ambiguous or unclear phrases: "What do you mean here? I don't understand what this word refers to." (Karanfil); b) "I am trying to interpret what this topic sentence wants to say; it is kind of confusing." (Mehmet)

2. Discern rhetorical structure: a) “I can see a thesis statement, background information, and two arguments.” (Arif); b) “This paragraph has every trait essential to an essay like, well, a topic sentence, a body part expanding the main idea, and a concluding sentence.” (Mesut).
3. Summarize ideas or propositions: a) “So the student talks about the language issue stating that foreign teachers who can speak Turkish should be hired.” (Cemil); b) “This paragraph states that non-native teachers cannot create a natural learning environment so native teachers are necessary for practicing English.” (Naime)

Rhetorical and Ideational Focus—Judgment Strategies

1. Assess reasoning, logic, or topic development: a) “The body paragraphs support the given thesis statement with logical relevant examples.” (Tugce); b) “The essay attempts to elaborate into the topic but cannot support the ideas with details and critical thought.” (Nazende)
2. Assess task completion and relevance: a) “He is giving information about something else. So, we can easily say that the essay is off-topic.” (Derya); b) “But, uh, the topic is still not related; it didn’t mention anything about what the topic required.” (Bilal); c) “I have read the first two paragraphs but I still don’t know what the topic is.” (Pamira)
3. Assess coherence: a) “Well, the essay promises to talk about both sides in the beginning but in the end I see something else. There is no coherence, I can say.” (Seren); b) “ Now, the student contradicts himself here.” (Efe)
4. Assess interest, originality, or creativity: a) “Yes, here comes another different point of view. I like this one too.” (Kamil); b) “It is a pretentious but attractive title.” (Yakup)
5. Identify redundancies: a) “There are some repetitions in the second paragraph.” (Karanfil); b) “It feels like this student is talking about the same thing over and over.” (Naime)

6. Assess text organization: a) “You start to wrap up the essay with ‘consequently’ but it’s not a new paragraph and seems to be a continuation of the body paragraph.” (Sertap); b) “Neither the whole essay nor each paragraph has been organized appropriately.” (Guney)
7. Assess style, register, or genre: a) “Phrases like ‘I think’ should not be used in the essay since they sound subjective.” (Kayra); b) “This student uses the phrase, ‘if you ask me,’ frequently. Why would I ask you? Are we talking?” (Pamira)
8. Rate ideas or rhetoric: a) “Actually, I like the ideas and the way they are presented.” (Remzi); b) “Now the ideas are clear and good in this paragraph but he could have followed a better rhetorical structure.” (Celal)

Language Focus—Interpretation Strategies

1. Observe layout: a) “Looking over the essay, uh, the student didn’t use the paragraph indents at all.” (Arif); b) “I do not like the formatting at first glance. The paragraphs are not indented and there is no space after the punctuation.” (Mesut)
2. Classify errors into types: a) “He is missing a preposition and there is a tense error here.” (Hamit); b) “I see punctuation mistakes and subject verb agreement problems.” (Sertap)
3. Edit phrases for interpretation: a) “He should have used ‘before’ in this sentence. ‘They have never known before.’” (Efe); b) “I think he meant to say ‘foreign people’ because ‘English people’ does not make sense.” (Seren)

Language Focus—Judgment Strategies

1. Assess quantity of written production: a) “I am looking over the text and I see an introduction, two body paragraphs and a conclusion. The conclusion seems a little short though.” (Naime); b) “This is a …one, two, three, and four-paragraph essay.” (Tugce)
2. Assess comprehensibility: a) “I didn’t understand what he wanted to say.” (Oznur); b) It took a long time to read the essay. It was not easy to understand it.” (Onur)

3. Consider gravity of errors: a) “It started well but later I saw some grammar mistakes, but they were not very important and did not interfere with the meaning.” (Oznur); b) “There are a lot of grammar mistakes and it affects the comprehensibility of the text.” (Adalya).
4. Consider error frequency: a) “This is the second time the student has struggled with a passive sentence.” (Bilal); b) “There are a lot of grammar, syntax, and spelling mistakes in this text.” (Kaan)
5. Assess fluency: (a) “The flow between sentences is not very good. The writer should try better to connect the sentences.” (Vahdet); b) “Nice transitions can be seen throughout the entire essay so that it has no issues with fluency at all.” (Derya)
6. Consider lexis: a) “The student used the word ‘ostracize.’” What does it mean? Let me underline it and I will check it later.” (Cemil); b) “This student has no control of appropriate word choice.” (Efe)
7. Consider syntax or morphology: “He uses passive forms and modals together without any problems.” (Hamit); b) “I see a syntax problem again, because he tries to write long sentences to fill the given word limit, I think.” (Kayra)
8. Consider spelling or punctuation: a) “There is no comma here; there should be a comma right here.” (Vahdet); b) “He should have used comma after the phrase in my opinion.” (Hasan); c) “I think there is a spelling mistake here.” (Adalya)
9. Rate language overall: a) “Well, the essay is so good that it feels like a native speaker of English wrote it.” (Ahmet); b) “Until here, this student uses good language and expresses himself well.” (Kamil)

(adapted from Cumming, Kantor, & Powers, 2002, pp. 93-94)

Appendix J: Assessment Instructions for Raters

Dear Rater,

I am currently working on my Ph.D. thesis at Çanakkale Onsekiz Mart University in the English Language Teaching Department. The purpose of this research study is to investigate the impact of rater experience and essay quality on rater behaviors and scoring. In this respect, you are kindly requested to assess the essays in the pack using the analytic rubric. There are 50 essays in total and 50 rubrics provided, one for each essay. You will use think-aloud protocol while assessing 16 of the essays and those essays have been tagged in the set. Please pay attention to the following items while evaluating the papers in the folder. I would like to thank you in advance for your valuable contribution. With my best regards.

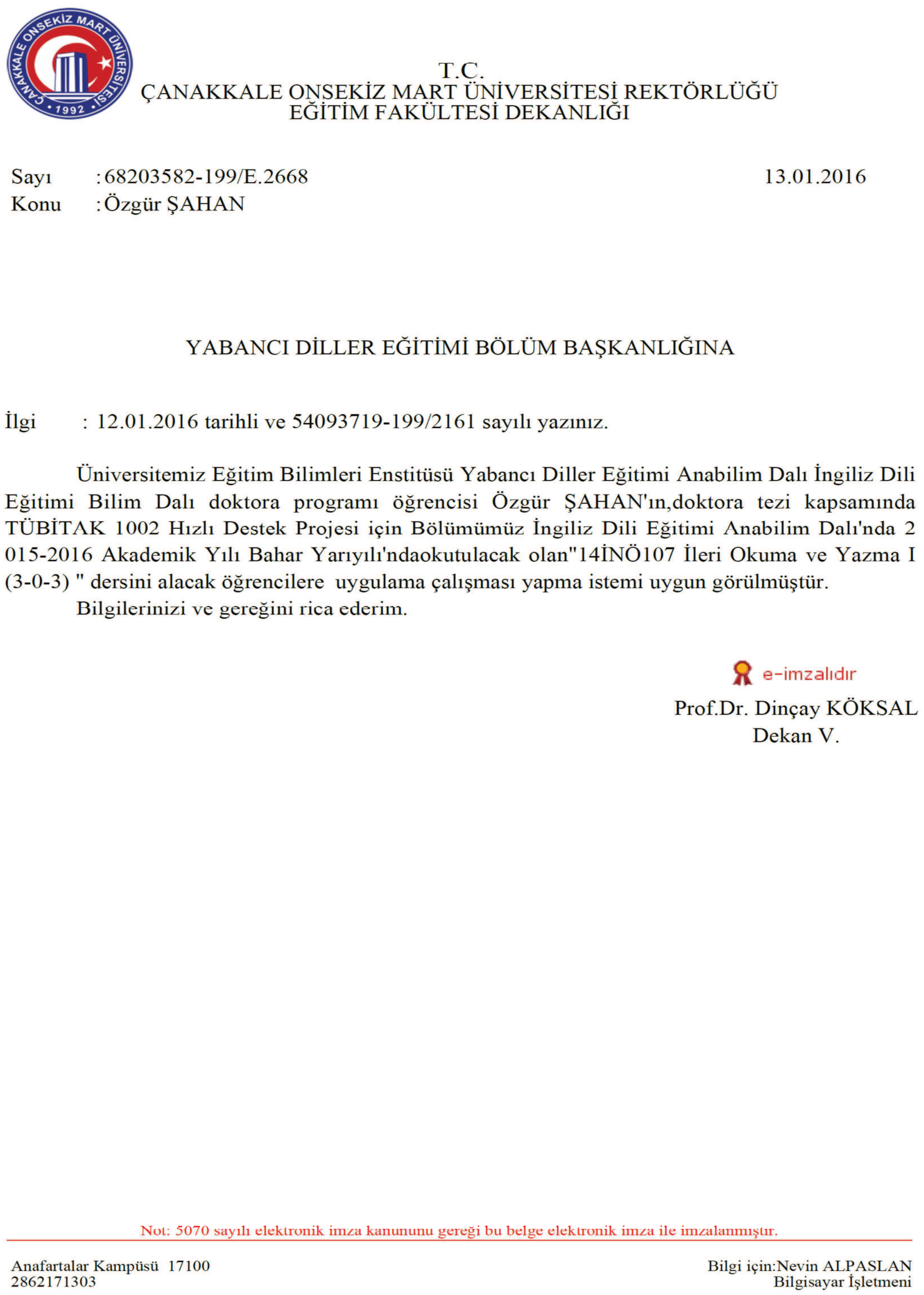
Özgür ŞAHAN
Assistant Director
Bursa Technical University
School of Foreign Languages
e-mail: ozgursahan66@hotmail.com

Essay Topic: Some people think that English teachers working at primary schools and high schools are insufficient to teach English effectively. Therefore, Ministry of Education in Turkey is thinking of hiring native English-speaking teachers to support English language education. Do you think that English teachers in Turkey are qualified enough for teaching English to the students or should English language education in Turkey be supported by native English-speaking teachers? Use specific reasons and examples to develop your essay.

- Please read the essay topic written above before assessing the papers.
- The essays were collected from first-year students enrolled in an ELT Department at a state university.
- Please be aware that the students' English language proficiency level is B1/B2.

- Students were expected to write 500- to 700-word essays.
- The essays were submitted using *Turnitin* to avoid plagiarism incidents.
- You are expected to assess the essays using the analytic rubric provided.
- Do not memorize the scoring rubric; instead, use the rubric for each essay.
- The rubric includes five components. Please write your scores for each component (e.g. grammar, content, etc.) on the far right column and the total score at the top of the page (your score).
- You can give partial points within the given maximum score range for each component (0.1, 1.3, 1.7, etc. out of 2.0)
- After assigning your score, please indicate three reasons that impact your decision most for the essay at the top of each paper.
- Do not negotiate your decision with anyone else.
- Please evaluate the essays individually rather than comparing them to other essays in the set.

Appendix K: Official Permission from Dean’s Office for Data Collection



T.C.
ÇANAKKALE ONSEKİZ MART ÜNİVERSİTESİ REKTÖRLÜĞÜ
EĞİTİM FAKÜLTESİ DEKANLIĞI

Sayı : 68203582-199/E.2668 13.01.2016
Konu : Özgür ŞAHAN

YABANCI DİLLER EĞİTİMİ BÖLÜM BAŞKANLIĞINA

İlgi : 12.01.2016 tarihli ve 54093719-199/2161 sayılı yazınız.

Üniversitemiz Eğitim Bilimleri Enstitüsü Yabancı Diller Eğitimi Anabilim Dalı İngiliz Dili Eğitimi Bilim Dalı doktora programı öğrencisi Özgür ŞAHAN'ın,doktora tezi kapsamında TÜBİTAK 1002 Hızlı Destek Projesi için Bölümümüz İngiliz Dili Eğitimi Anabilim Dalı'nda 2 015-2016 Akademik Yılı Bahar Yarıyılı'ndaokutulacak olan"14İNÖ107 İleri Okuma ve Yazma I (3-0-3) " dersini alacak öğrencilere uygulama çalışması yapma istemi uygun görülmüştür.

Bilgilerinizi ve gereğini rica ederim.

e-imzalıdır
Prof.Dr. Dinçay KÖKSAL
Dekan V.

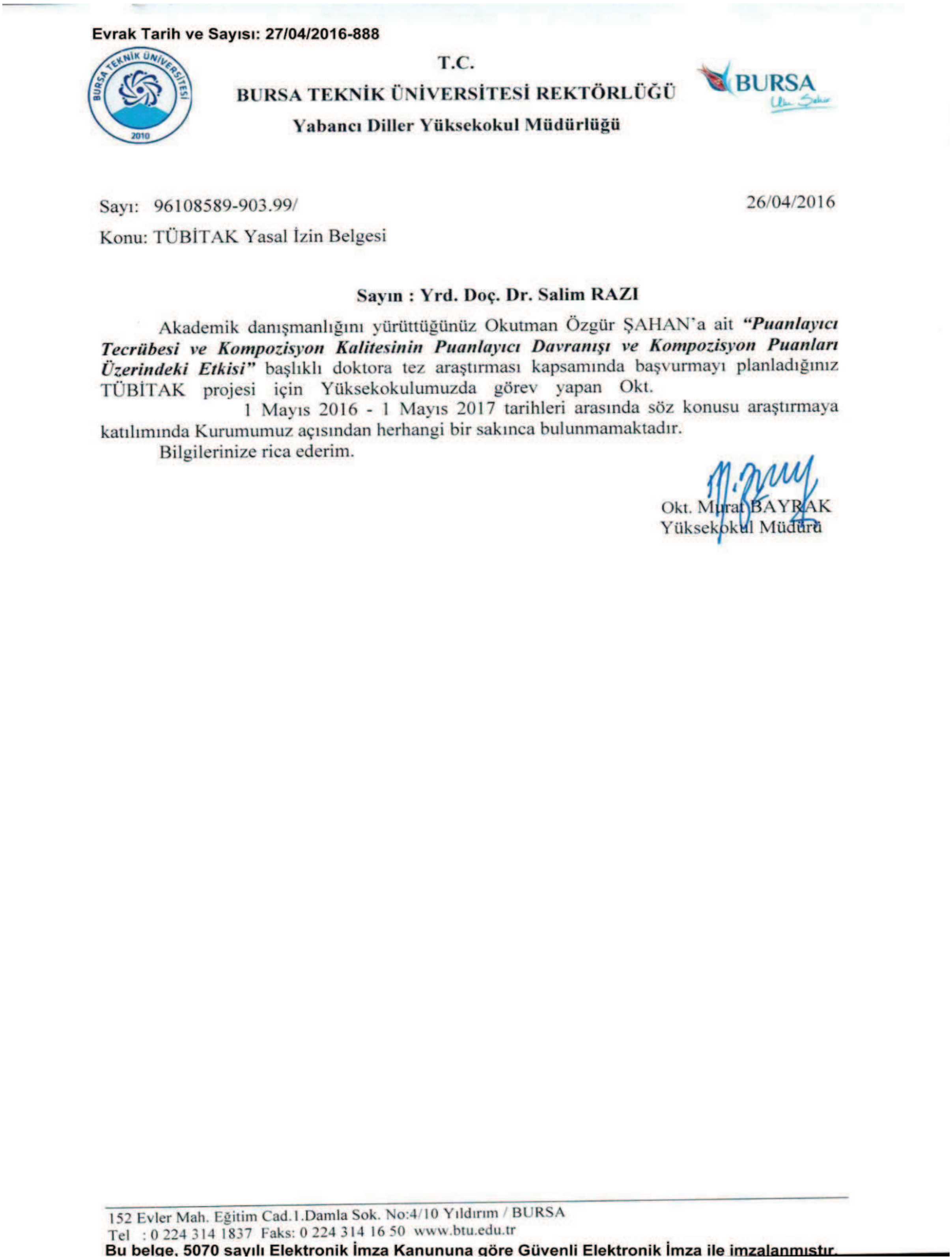
Not: 5070 sayılı elektronik imza kanununu gereği bu belge elektronik imza ile imzalanmıştır.

Anafartalar Kampüsü 17100
2862171303

Bilgi için:Nevin ALPASLAN
Bilgisayar İşletmeni

Appendix L: Official Permission Forms for Raters' Research Participation

The following signed permission forms were obtained as part of a research grant application to the Scientific and Technological Research Council of Turkey (abbreviated in Turkish as TÜBİTAK). Although the TÜBİTAK application was unsuccessful, the researcher has included the official permission forms for the sake of ethical transparency. The participants' names have been removed from the documents.



Evrak Tarih ve Sayısı: 27/04/2016-888

T.C.

BURSA TEKNİK ÜNİVERSİTESİ REKTÖRLÜĞÜ

Yabancı Diller Yüksekokul Müdürlüğü

BURSA

Sayı: 96108589-903.99/ 26/04/2016

Konu: TÜBİTAK Yasal İzin Belgesi

Sayın : Yrd. Doç. Dr. Salim RAZI

Akademik danışmanlığını yürüttüğünüz Okutman Özgür ŞAHAN'a ait ***"Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi"*** başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için Yüksekokulumuzda görev yapan Okt.

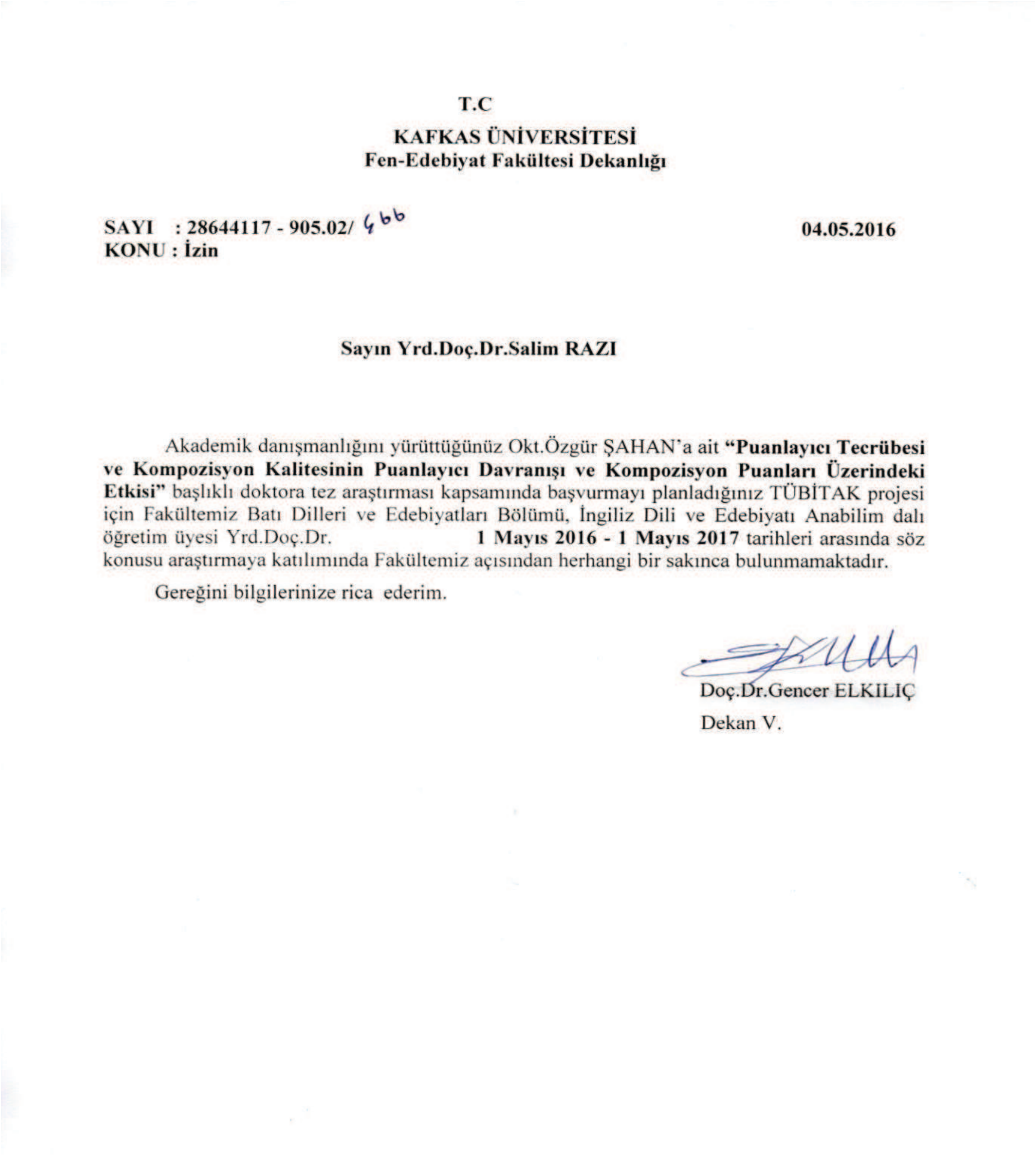
1 Mayıs 2016 - 1 Mayıs 2017 tarihleri arasında söz konusu araştırmaya katılımında Kurumumuz açısından herhangi bir sakınca bulunmamaktadır.

Bilgilerinize rica ederim.

Okt. Murat BAYRAK
Yüksekokul Müdürü

152 Evler Mah. Eğitim Cad.1.Damla Sok. No:4/10 Yıldırım / BURSA
Tel : 0 224 314 1837 Faks: 0 224 314 16 50 www.btu.edu.tr

Bu belge, 5070 sayılı Elektronik İmza Kanununa göre Güvenli Elektronik İmza ile imzalanmıştır.



T.C

KAFKAS ÜNİVERSİTESİ

Fen-Edebiyat Fakültesi Dekanlığı

SAYI : 28644117 - 905.02/ 466 **04.05.2016**

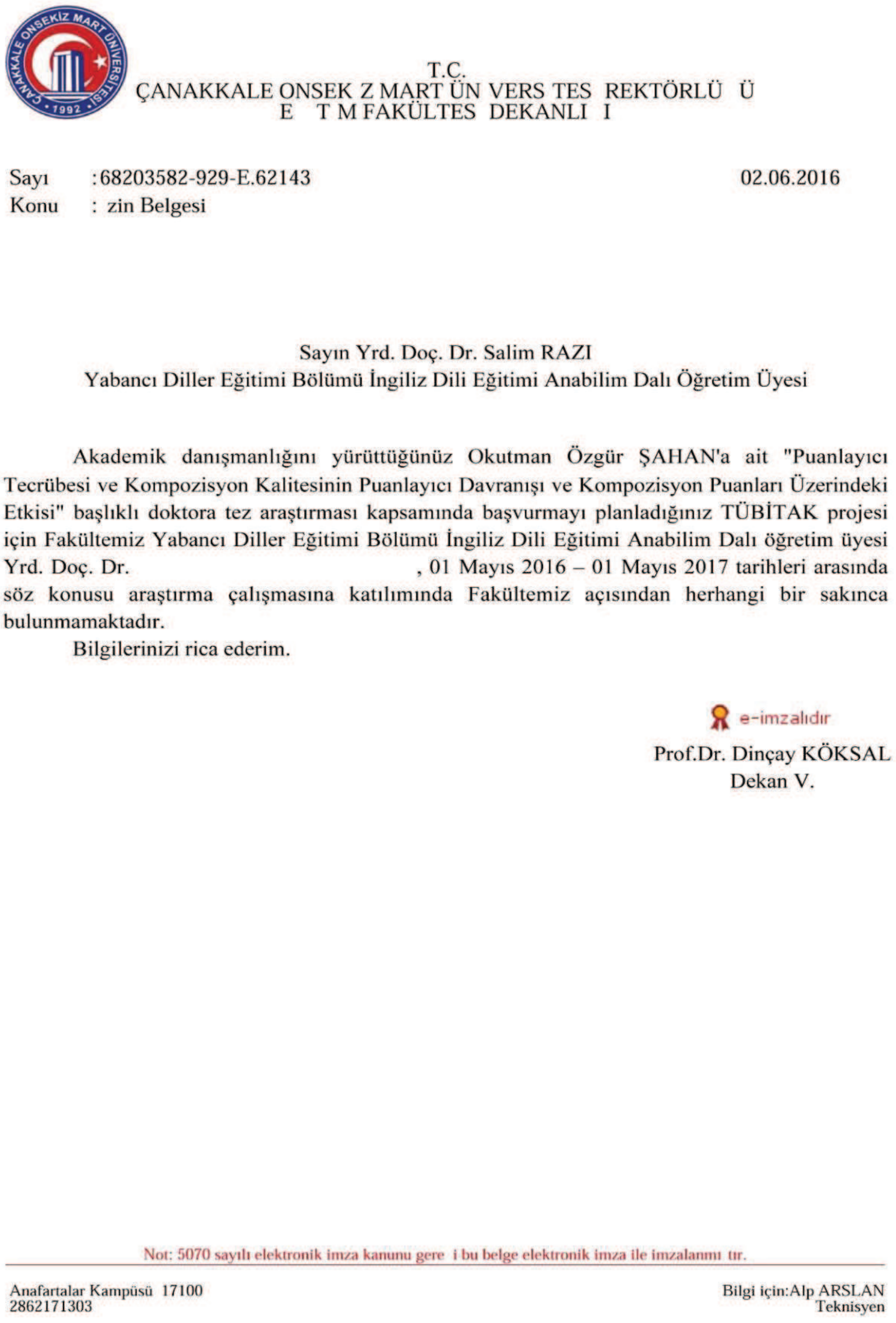
KONU : İzin

Sayın Yrd.Doç.Dr.Salim RAZI

Akademik danışmanlığını yürüttüğünüz Okt.Özgür ŞAHAN'a ait **"Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi"** başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için Fakültemiz Batı Dilleri ve Edebiyatları Bölümü, İngiliz Dili ve Edebiyatı Anabilim dalı öğretim üyesi Yrd.Doç.Dr. **1 Mayıs 2016 - 1 Mayıs 2017** tarihleri arasında söz konusu araştırmaya katılımında Fakültemiz açısından herhangi bir sakınca bulunmamaktadır.

Gereğini bilgilerinize rica ederim.

Doç.Dr.Gencer ELKILIÇ
Dekan V.



T.C.
ÇANAKKALE ONSEK Z MART ÜN VERS TES REKTÖRLÜ Ü
E T M FAKÜLTES DEKANLI I

Sayı : 68203582-929-E.62143 02.06.2016
Konu : zin Belgesi

Sayın Yrd. Doç. Dr. Salim RAZI
Yabancı Diller Eğitimi Bölümü İngiliz Dili Eğitimi Anabilim Dalı Öğretim Üyesi

Akademik danışmanlığını yürüttüğünüz Okutman Özgür ŞAHAN'a ait "Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi" başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için Fakültemiz Yabancı Diller Eğitimi Bölümü İngiliz Dili Eğitimi Anabilim Dalı öğretim üyesi Yrd. Doç. Dr. , 01 Mayıs 2016 – 01 Mayıs 2017 tarihleri arasında söz konusu araştırma çalışmasına katılımında Fakültemiz açısından herhangi bir sakınca bulunmamaktadır.

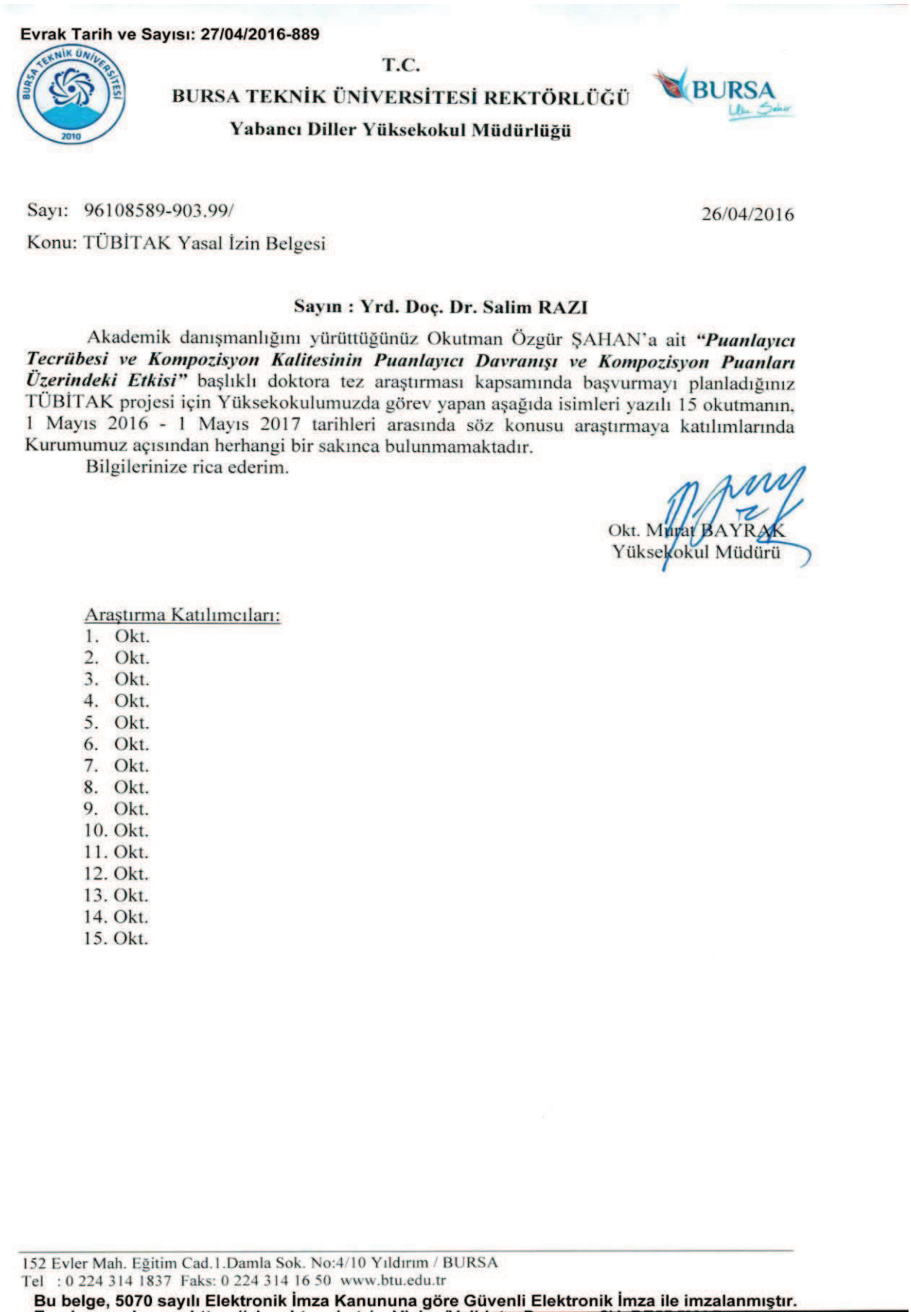
Bilgilerinizi rica ederim.

e-imzalıdır
Prof.Dr. Dinçay KÖKSAL
Dekan V.

Not: 5070 sayılı elektronik imza kanunu gere i bu belge elektronik imza ile imzalanmı tır.

Anafartalar Kampüsü 17100
2862171303

Bilgi için:Alp ARSLAN
Teknisyen



Evrak Tarih ve Sayısı: 27/04/2016-889

T.C.
BURSA TEKNİK ÜNİVERSİTESİ REKTÖRLÜĞÜ
Yabancı Diller Yüksekokul Müdürlüğü

Sayı: 96108589-903.99/ 26/04/2016

Konu: TÜBİTAK Yasal İzin Belgesi

Sayın : Yrd. Doç. Dr. Salim RAZI

Akademik danışmanlığını yürüttüğünüz Okutman Özgür ŞAHAN'a ait ***"Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi"*** başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için Yüksekokulumuzda görev yapan aşağıda isimleri yazılı 15 okutmanın, 1 Mayıs 2016 - 1 Mayıs 2017 tarihleri arasında söz konusu araştırmaya katılımlarında Kurumumuz açısından herhangi bir sakınca bulunmamaktadır.

Bilgilerinize rica ederim.

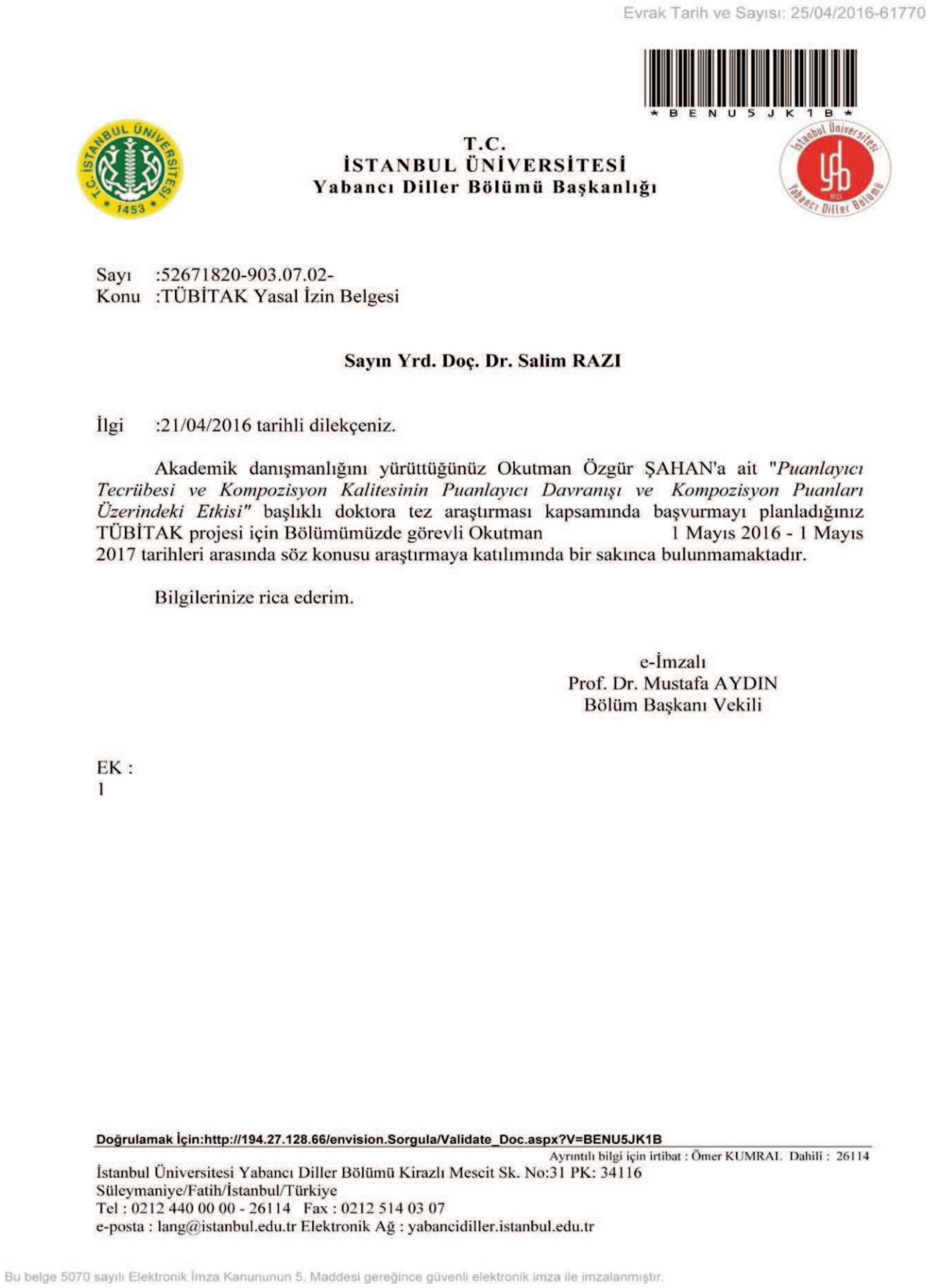
Okt. Murat BAYRAK
Yüksekokul Müdürü

Araştırma Katılımcıları:

1. Okt.
2. Okt.
3. Okt.
4. Okt.
5. Okt.
6. Okt.
7. Okt.
8. Okt.
9. Okt.
10. Okt.
11. Okt.
12. Okt.
13. Okt.
14. Okt.
15. Okt.

152 Evler Mah. Eğitim Cad.1.Damla Sok. No:4/10 Yıldırım / BURSA
Tel : 0 224 314 1837 Faks: 0 224 314 16 50 www.btu.edu.tr

Bu belge, 5070 sayılı Elektronik İmza Kanununa göre Güvenli Elektronik İmza ile imzalanmıştır.



Evrak Tarih ve Sayısı: 25/04/2016-61770

* B E N U 5 J K 1 B *

T.C.
İSTANBUL ÜNİVERSİTESİ
Yabancı Diller Bölümü Başkanlığı

Sayı :52671820-903.07.02-
Konu :TÜBİTAK Yasal İzin Belgesi

Sayın Yrd. Doç. Dr. Salim RAZI

İlgi :21/04/2016 tarihli dilekçeniz.

Akademik danışmanlığını yürüttüğünüz Okutman Özgür ŞAHAN'a ait "*Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi*" başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için Bölümümüzde görevli Okutman 1 Mayıs 2016 - 1 Mayıs 2017 tarihleri arasında söz konusu araştırmaya katılımında bir sakınca bulunmamaktadır.

Bilgilerinize rica ederim.

e-İmzalı
Prof. Dr. Mustafa AYDIN
Bölüm Başkanı Vekili

EK :
1

Doğrulamak İçin:http://194.27.128.66/envision.Sorgula/Validate_Doc.aspx?V=BENU5JK1B

Ayrıntılı bilgi için irtibat : Ömer KUMRAL Dahili : 26114

İstanbul Üniversitesi Yabancı Diller Bölümü Kirazlı Mescit Sk. No:31 PK: 34116
Süleymaniye/Fatih/İstanbul/Türkiye
Tel : 0212 440 00 00 - 26114 Fax : 0212 514 03 07
e-posta : lang@istanbul.edu.tr Elektronik Ağ : yabancidiller.istanbul.edu.tr

Bu belge 5070 sayılı Elektronik İmza Kanununun 5. Maddesi gereğince güvenli elektronik imza ile imzalanmıştır.

T. C.
TRAKYA ÜNİVERSİTESİ REKTÖRLÜĞÜ
Zorunlu Ortak Servis Dersleri Koordinatörlüğü

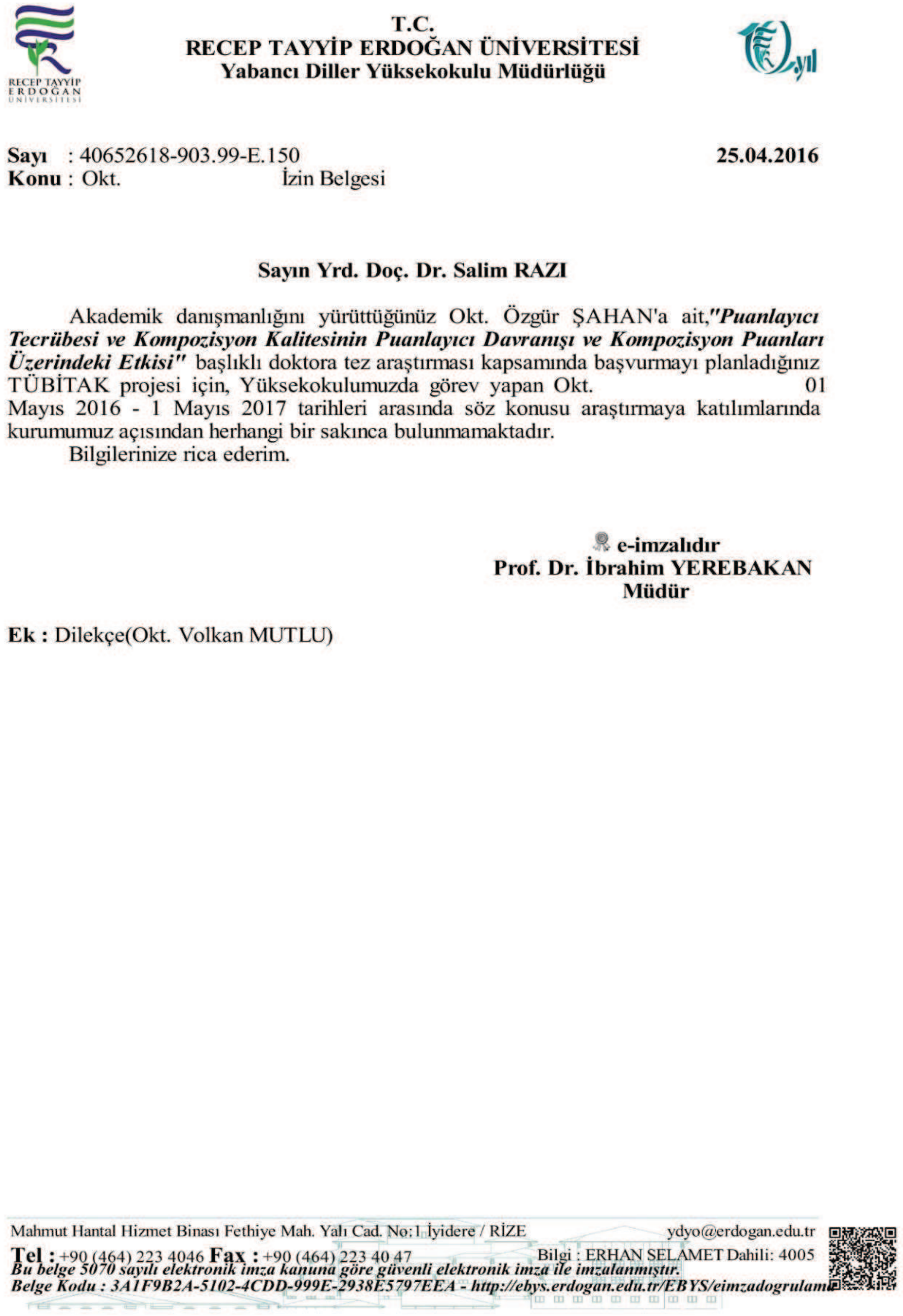
Sayı: 20916276-903-539 **09/05/2016**

Sayın: Yrd. Doç. Dr. Salim RAZI

Akademik danışmanlığını yürüttüğünüz Okutman Özgür ŞAHAN 'a ait **'Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi'** başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için Zorunlu Ortak Servis Derslerinde görev yapan Okt.
1 Mayıs 2016 - 1 Mayıs 2017 tarihleri arasında söz konusu araştırmaya katılımında Kurumumuz açısından herhangi bir sakınca bulunmamaktadır.

Bilgilerinize rica ederim.

Prof. Dr. Mustafa ÖZCAN
Zorunlu Ortak Servis Dersleri
Koordinatörü



T.C.
RECEP TAYYİP ERDOĞAN ÜNİVERSİTESİ
Yabancı Diller Yüksekokulu Müdürlüğü

Sayı : 40652618-903.99-E.150 **25.04.2016**
Konu : Okt. İzin Belgesi

Sayın Yrd. Doç. Dr. Salim RAZI

Akademik danışmanlığını yürüttüğünüz Okt. Özgür ŞAHAN'a ait,***"Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi"*** başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için, Yüksekokulumuzda görev yapan Okt. 01 Mayıs 2016 - 1 Mayıs 2017 tarihleri arasında söz konusu araştırmaya katılımlarında kurumumuz açısından herhangi bir sakınca bulunmamaktadır.

Bilgilerinize rica ederim.

e-imzalıdır
Prof. Dr. İbrahim YEREBAKAN
Müdür

Ek : Dilekçe(Okt. Volkan MUTLU)

Mahmut Hantal Hizmet Binası Fethiye Mah. Yalı Cad. No:1 İyidere / RİZE ydyo@erdogan.edu.tr
Tel : +90 (464) 223 4046 **Fax :** +90 (464) 223 40 47 Bilgi : ERHAN SELAMET Dahili: 4005
Bu belge 5070 sayılı elektronik imza kanuna göre güvenli elektronik imza ile imzalanmıştır.
Belge Kodu : 3A1F9B2A-5102-4CDD-999E-2938E5797EEA - http://ebys.erdogan.edu.tr/EBYS/eimzadogrulama

BEAM4LKV6

T.C.
PAMUKKALE ÜNİVERSİTESİ
Yabancı Diller Yüksekokulu

Sayı :63788039 -900/12929 18/07/2016
Konu :TÜBİTAK Çalışma İzni

Sayın Yrd.Doç.Dr. Salim RAZI

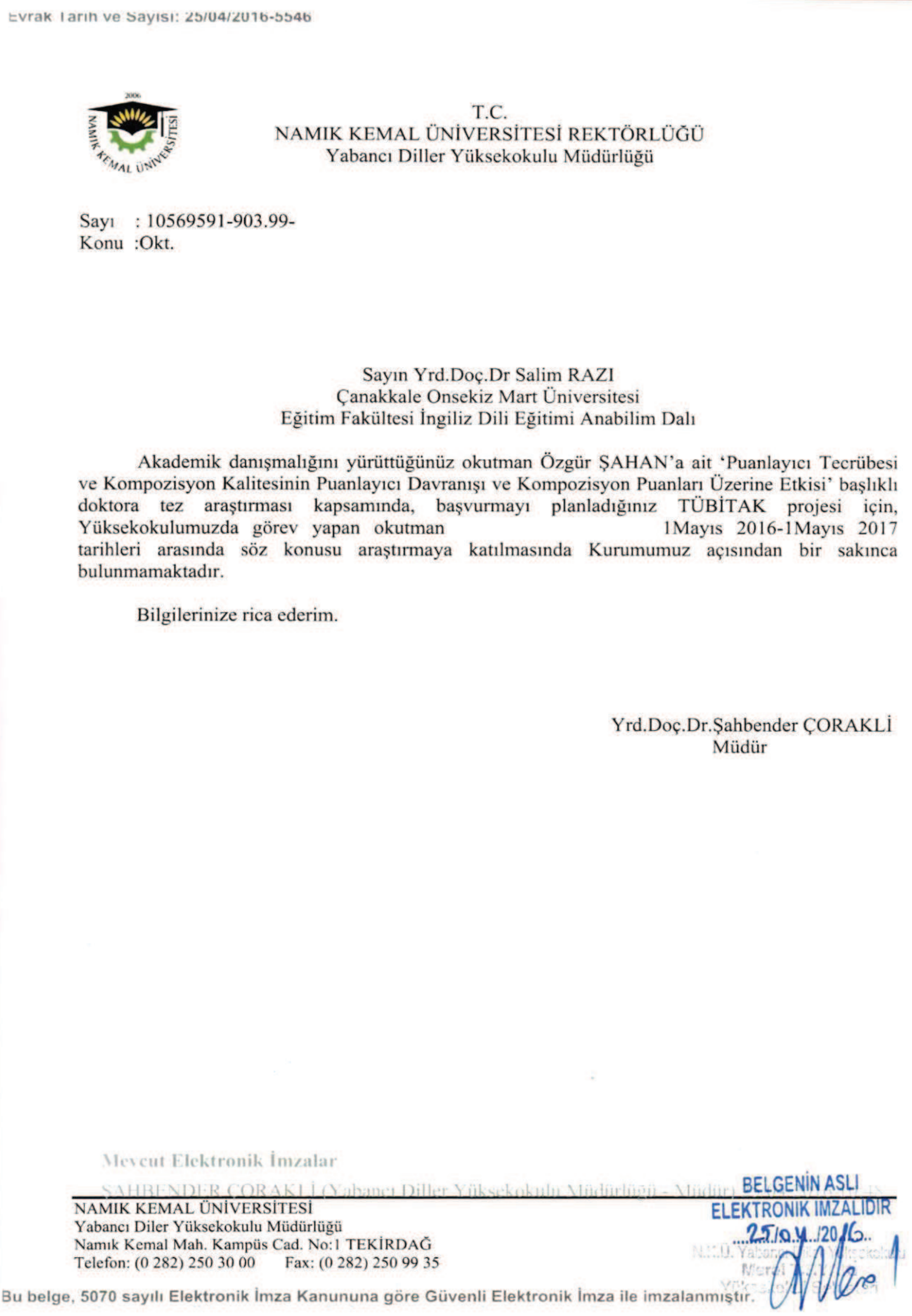
Akademik danışmanlığı yürüttüğünüz Okt. Özgür ŞAHAN'a ait ***"Puanlayıcı Tücrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi"*** başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için Yüksekokulumuz öğretim elemanı Okt.
TUNA'nın 01 Mayıs 2016 - 01 Mayıs 2017 tarihleri arasında söz konusu araştırmaya katılımında Birimimiz açısından herhangi bir sakınca bulunmamaktadır.

Bilgilerinizi rica ederim.

Prof. Dr. Turgut TOK
Müdür

12/07/2016 Mem. : H.TATLICI

Evrakı Doğrulamak İçin : http://dys.pau.edu.tr/enVision/Dogrula/AM4LKV6
Pamukkale Üniversitesi Yabancı Diller Yüksekokulu 20160 Kınıklı Yerleşkesi DENİZLİ
Ayrıntılı bilgi için irtibat : Hülya TATLICI
Tel: 0 (258) 296 15 06 Faks: 0 (258) 296 15 64
E-Posta: ydyo@pau.edu.tr Elektronik Ağ:http://www.pau.edu.tr/ydyo



Evrak Tarih ve Sayısı: 25/04/2016-5546

T.C.
NAMIK KEMAL ÜNİVERSİTESİ REKTÖRLÜĞÜ
Yabancı Diller Yüksekokulu Müdürlüğü

Sayı : 10569591-903.99-
Konu :Okt.

Sayın Yrd.Doç.Dr Salim RAZI
Çanakkale Onsekiz Mart Üniversitesi
Eğitim Fakültesi İngiliz Dili Eğitimi Anabilim Dalı

Akademik danışmalığını yürüttüğünüz okutman Özgür ŞAHAN'a ait 'Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerine Etkisi' başlıklı doktora tez araştırması kapsamında, başvurmayı planladığınız TÜBİTAK projesi için, Yüksekokulumuzda görev yapan okutman 1Mayıs 2016-1Mayıs 2017 tarihleri arasında söz konusu araştırmaya katılmasında Kurumumuz açısından bir sakınca bulunmamaktadır.

Bilgilerinize rica ederim.

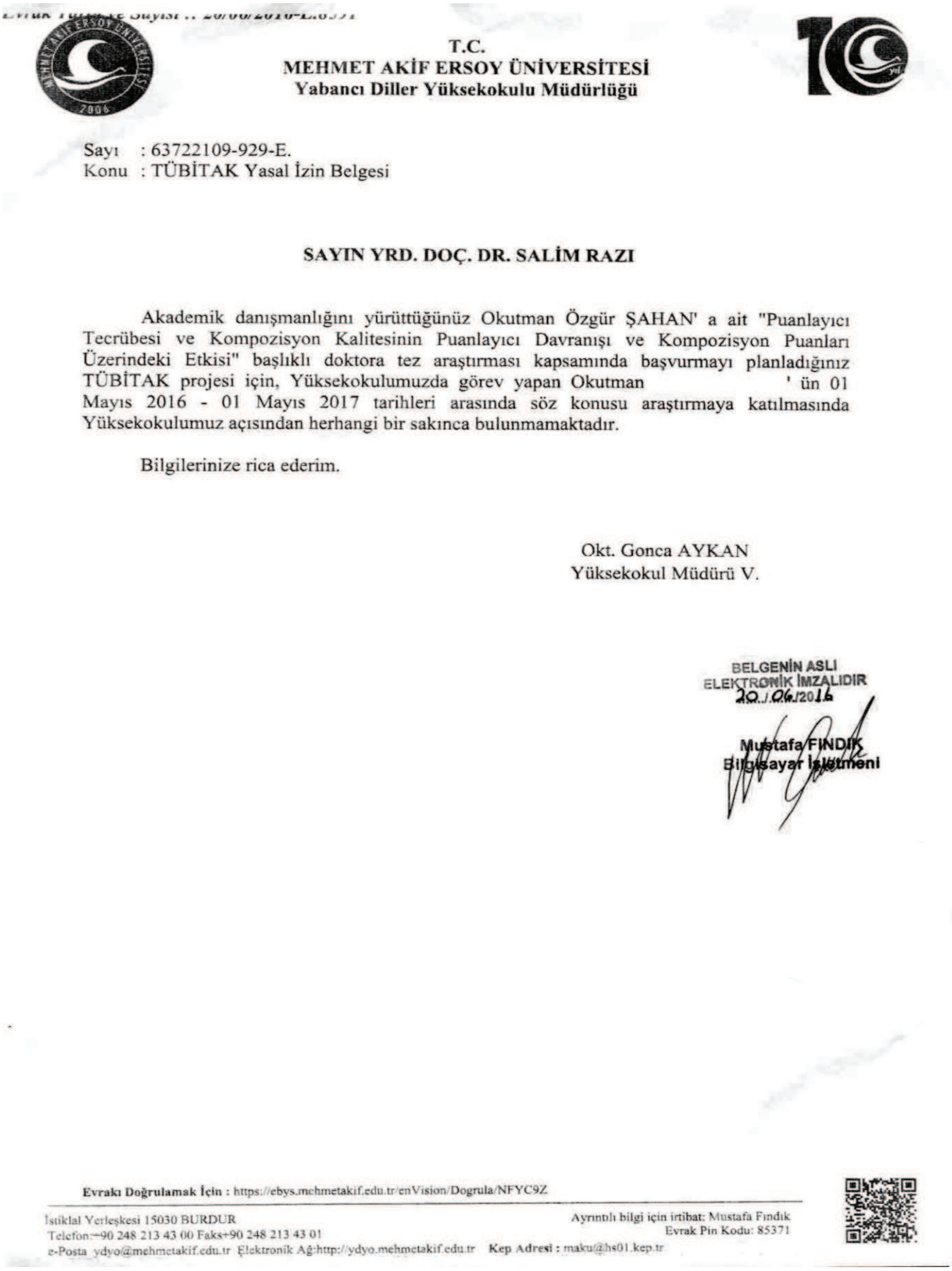
Yrd.Doç.Dr.Şahbender ÇORAKLİ
Müdür

Mevcut Elektronik İmzalar
ŞAHBENDER ÇORAKLİ (Yabancı Diller Yüksekokulu Müdürlüğü - Müdür)

NAMIK KEMAL ÜNİVERSİTESİ
Yabancı Diler Yüksekokulu Müdürlüğü
Namık Kemal Mah. Kampüs Cad. No:1 TEKİRDAĞ
Telefon: (0 282) 250 30 00 Fax: (0 282) 250 99 35

BELGENİN ASLI
ELEKTRONIK IMZALIDIR
...25/04../2016..
N.K.Ü. Yabancı Diller Yüksekokulu

Bu belge, 5070 sayılı Elektronik İmza Kanununa göre Güvenli Elektronik İmza ile imzalanmıştır.



Evrak Tarih ve Sayısı : 20/06/2016-E.8391

T.C.
MEHMET AKİF ERSOY ÜNİVERSİTESİ
Yabancı Diller Yüksekokulu Müdürlüğü

Sayı : 63722109-929-E.
Konu : TÜBİTAK Yasal İzin Belgesi

SAYIN YRD. DOÇ. DR. SALİM RAZI

Akademik danışmanlığını yürüttüğünüz Okutman Özgür ŞAHAN' a ait "Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi" başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için, Yüksekokulumuzda görev yapan Okutman ' ün 01 Mayıs 2016 - 01 Mayıs 2017 tarihleri arasında söz konusu araştırmaya katılmasında Yüksekokulumuz açısından herhangi bir sakınca bulunmamaktadır.

Bilgilerinize rica ederim.

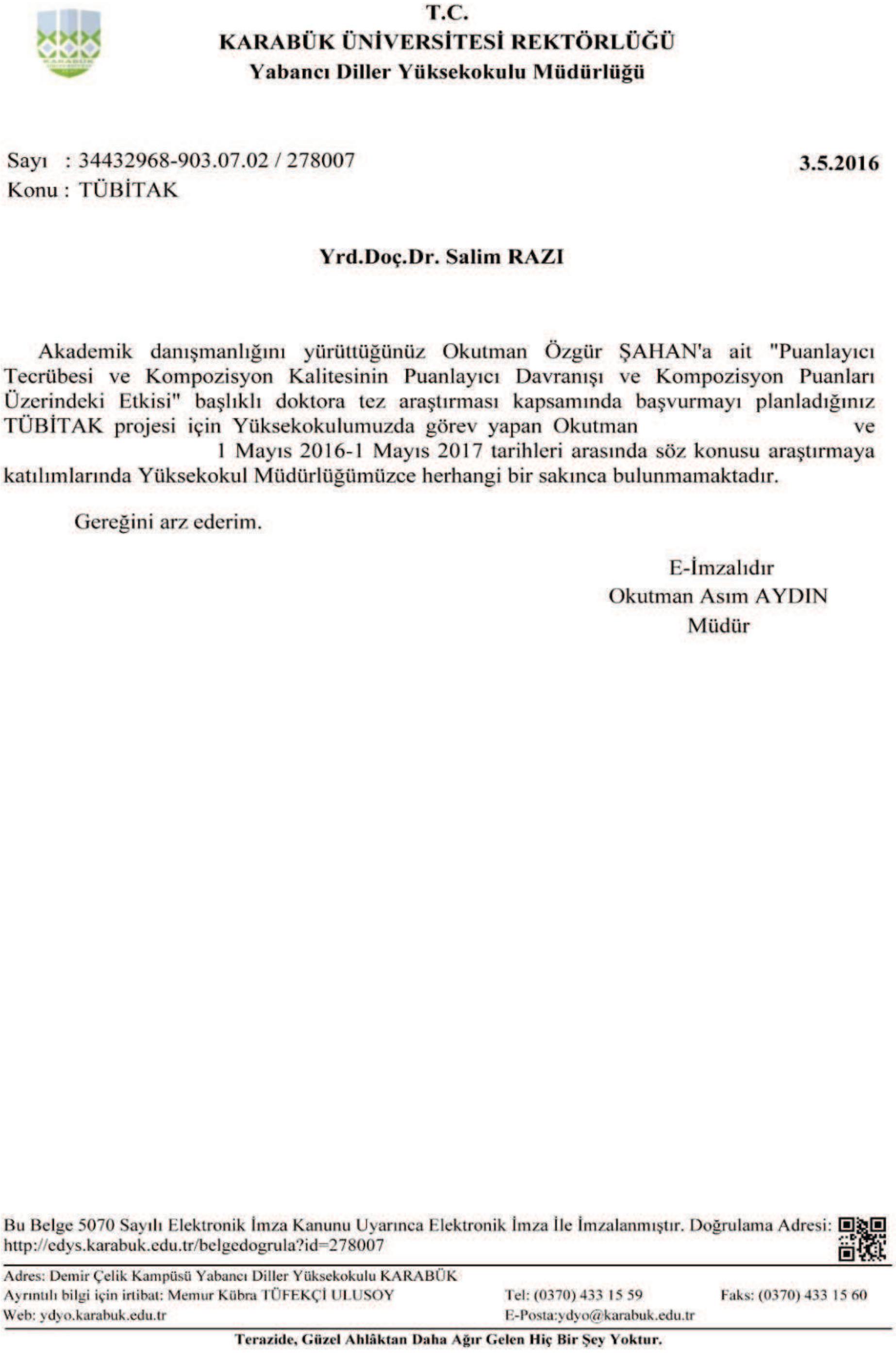
Okt. Gonca AYKAN
Yüksekokul Müdürü V.

BELGENİN ASLI
ELEKTRONİK İMZALIDIR
20./06./2016
Mustafa FINDIK
Bilgisayar İşletmeni

Evrakı Doğrulamak İçin : https://ebys.mehmetakif.edu.tr/enVision/Dogrula/NFYC9Z

İstiklal Yerleşkesi 15030 BURDUR
Telefon:+90 248 213 43 00 Faks+90 248 213 43 01
e-Posta ydyo@mehmetakif.edu.tr Elektronik Ağ:http://ydyo.mehmetakif.edu.tr **Kep Adresi :** maku@hs01.kep.tr

Ayrıntılı bilgi için irtibat: Mustafa Fındık
Evrak Pin Kodu: 85371



T.C.
KARABÜK ÜNİVERSİTESİ REKTÖRLÜĞÜ
Yabancı Diller Yüksekokulu Müdürlüğü

Sayı : 34432968-903.07.02 / 278007 **3.5.2016**
Konu : TÜBİTAK

Yrd.Doç.Dr. Salim RAZI

Akademik danışmanlığını yürüttüğünüz Okutman Özgür ŞAHAN'a ait "Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi" başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için Yüksekokulumuzda görev yapan Okutman ve 1 Mayıs 2016-1 Mayıs 2017 tarihleri arasında söz konusu araştırmaya katılımlarında Yüksekokul Müdürlüğümüzce herhangi bir sakınca bulunmamaktadır.

Gereğini arz ederim.

E-İmzalıdır
Okutman Asım AYDIN
Müdür

Bu Belge 5070 Sayılı Elektronik İmza Kanunu Uyarınca Elektronik İmza İle İmzalanmıştır. Doğrulama Adresi: http://edys.karabuk.edu.tr/belgedogrula?id=278007

Adres: Demir Çelik Kampüsü Yabancı Diller Yüksekokulu KARABÜK
Ayrıntılı bilgi için irtibat: Memur Kübra TÜFEKÇİ ULUSOY Tel: (0370) 433 15 59 Faks: (0370) 433 15 60
Web: ydyo.karabuk.edu.tr E-Posta:ydyo@karabuk.edu.tr

Terazide, Güzel Ahlâktan Daha Ağır Gelen Hiç Bir Şey Yoktur.

T.C.
HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜ MÜDÜRLÜĞÜ

Sayı: 51944218-010.99/1264 26/05/2016
Konu:
(İzin Belgesi)

Sayın Yrd. Doç. Dr. Salim RAZI

Akademik danışmanlığını yürüttüğünüz Okt.Özgür ŞAHAN'a ait "Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi" başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TUBİTAK projesi için, Enstitümüz Yabancı Diller Eğitimi Anabilim Dalında görev yapan Arş.Gör. 01 Mayıs 2016 – 01 Mayıs 2017 tarihleri arasında söz konusu araştırmaya katılımında herhangi bir sakınca bulunmamaktadır.

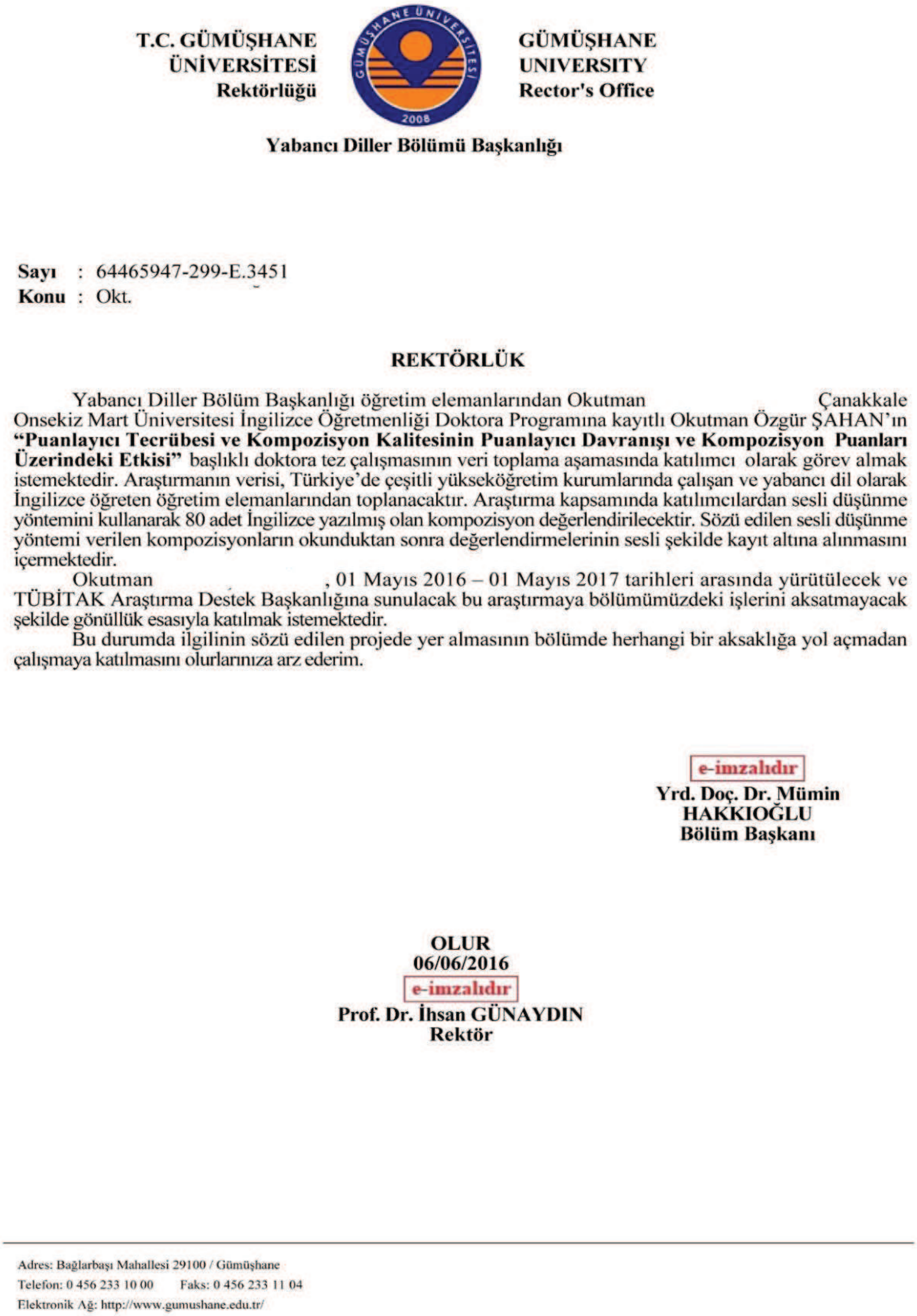
Bu belge ilgilinin talebi üzerine düzenlenmiştir.

Bilgilerinize rica ederim.

Prof.Dr. BERRİN AKMAN
Enstitü Müdürü

EKLER :
1 Dilekçe ( )

6115164767



T.C. GÜMÜŞHANE
ÜNİVERSİTESİ
Rektörlüğü

GÜMÜŞHANE
UNIVERSITY
Rector's Office

Yabancı Diller Bölümü Başkanlığı

Sayı : 64465947-299-E.3451
Konu : Okt.

REKTÖRLÜK

Yabancı Diller Bölüm Başkanlığı öğretim elemanlarından Okutman Çanakkale Onsekiz Mart Üniversitesi İngilizce Öğretmenliği Doktora Programına kayıtlı Okutman Özgür ŞAHAN'ın **"Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi"** başlıklı doktora tez çalışmasının veri toplama aşamasında katılımcı olarak görev almak istemektedir. Araştırmanın verisi, Türkiye'de çeşitli yükseköğretim kurumlarında çalışan ve yabancı dil olarak İngilizce öğreten öğretim elemanlarından toplanacaktır. Araştırma kapsamında katılımcılardan sesli düşünme yöntemini kullanarak 80 adet İngilizce yazılmış olan kompozisyon değerlendirilecektir. Sözü edilen sesli düşünme yöntemi verilen kompozisyonların okunduktan sonra değerlendirmelerinin sesli şekilde kayıt altına alınmasını içermektedir.

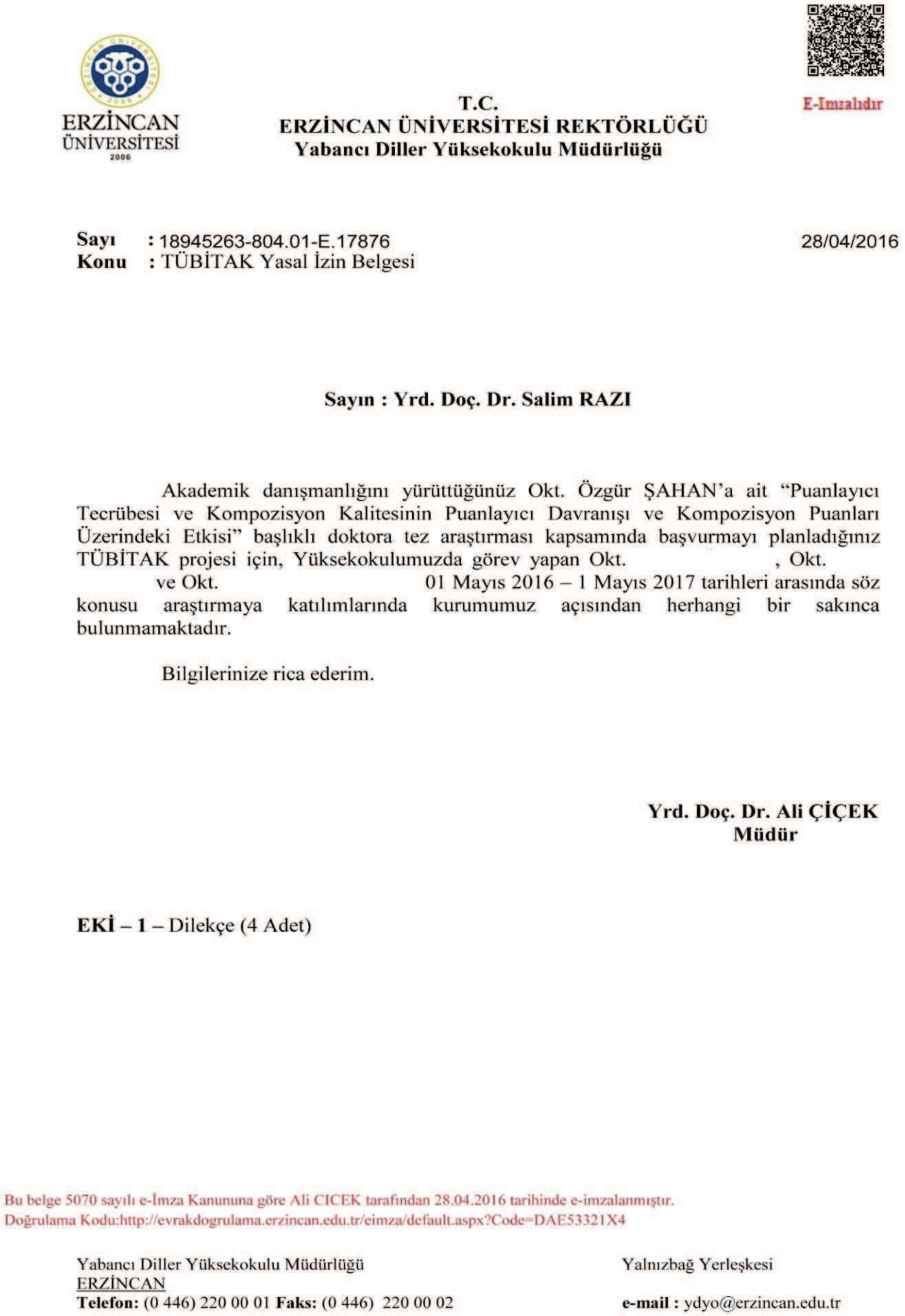
Okutman , 01 Mayıs 2016 – 01 Mayıs 2017 tarihleri arasında yürütülecek ve TÜBİTAK Araştırma Destek Başkanlığına sunulacak bu araştırmaya bölümümüzdeki işlerini aksatmayacak şekilde gönüllük esasıyla katılmak istemektedir.

Bu durumda ilgilinin sözü edilen projede yer almasının bölümde herhangi bir aksaklığa yol açmadan çalışmaya katılmasını olurlarınıza arz ederim.

e-imzalıdır
Yrd. Doç. Dr. Mümin HAKKIOĞLU
Bölüm Başkanı

OLUR
06/06/2016
e-imzalıdır
Prof. Dr. İhsan GÜNAYDIN
Rektör

Adres: Bağlarbaşı Mahallesi 29100 / Gümüşhane
Telefon: 0 456 233 10 00 Faks: 0 456 233 11 04
Elektronik Ağ: http://www.gumushane.edu.tr/



E-İmzalıdır

ERZİNCAN ÜNİVERSİTESİ 2006

T.C.
ERZİNCAN ÜNİVERSİTESİ REKTÖRLÜĞÜ
Yabancı Diller Yüksekokulu Müdürlüğü

Sayı : 18945263-804.01-E.17876 28/04/2016
Konu : TÜBİTAK Yasal İzin Belgesi

Sayın : Yrd. Doç. Dr. Salim RAZI

Akademik danışmanlığını yürüttüğünüz Okt. Özgür ŞAHAN'a ait "Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi" başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için, Yüksekokulumuzda görev yapan Okt. , Okt. ve Okt. 01 Mayıs 2016 – 1 Mayıs 2017 tarihleri arasında söz konusu araştırmaya katılımlarında kurumumuz açısından herhangi bir sakınca bulunmamaktadır.

Bilgilerinize rica ederim.

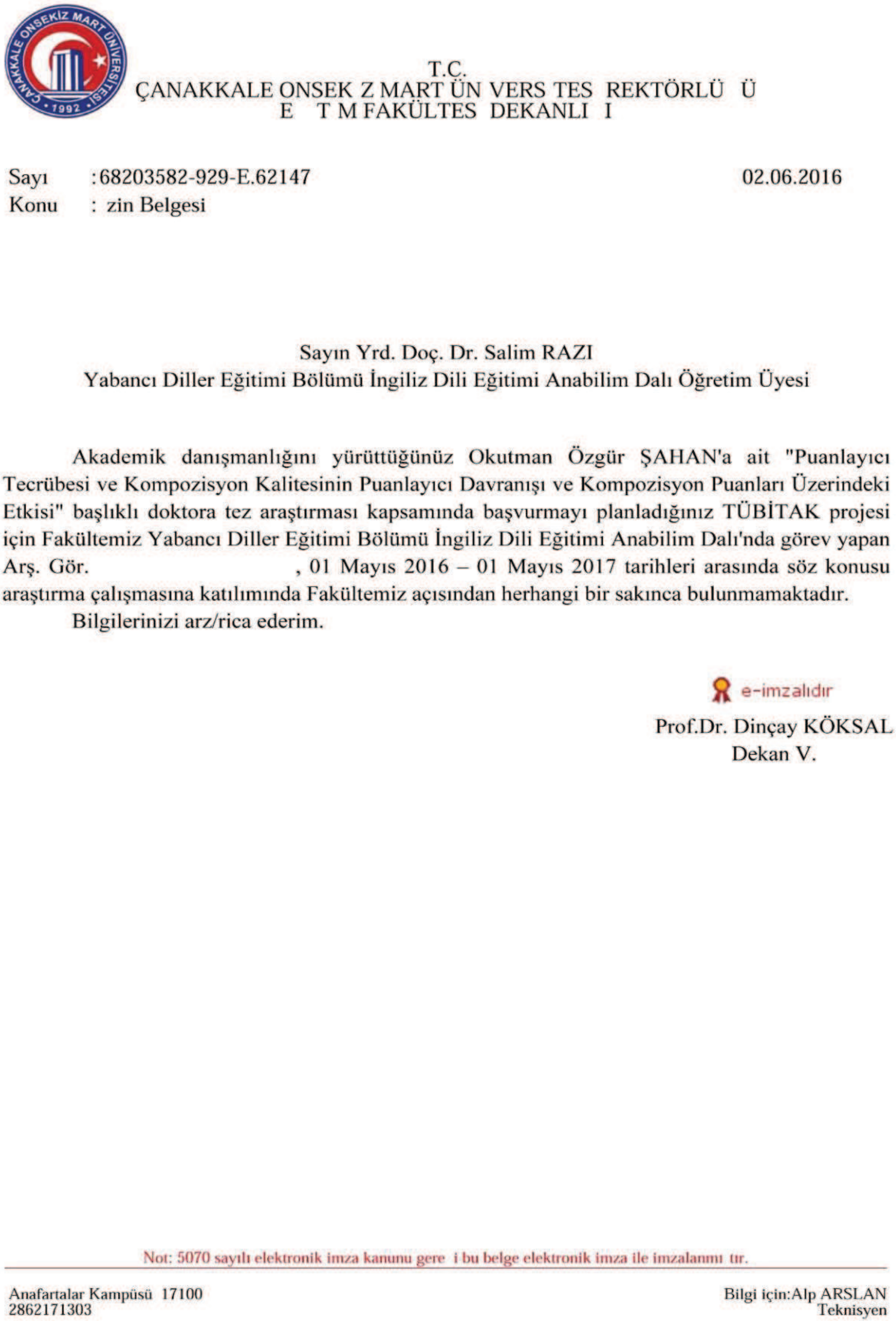
Yrd. Doç. Dr. Ali ÇİÇEK
Müdür

EKİ – 1 – Dilekçe (4 Adet)

Bu belge 5070 sayılı e-İmza Kanununa göre Ali CICEK tarafından 28.04.2016 tarihinde e-imzalanmıştır.
Doğrulama Kodu:http://evrakdogrulama.erzincan.edu.tr/eimza/default.aspx?Code=DAE53321X4

Yabancı Diller Yüksekokulu Müdürlüğü
ERZİNCAN
Telefon: (0 446) 220 00 01 **Faks:** (0 446) 220 00 02

Yalnızbağ Yerleşkesi
e-mail : ydyo@erzincan.edu.tr



T.C.
ÇANAKKALE ONSEK Z MART ÜN VERS TES REKTÖRLÜ Ü
E   T M FAKÜLTES DEKANLI I

Sayı :68203582-929-E.62147 02.06.2016
Konu : zin Belgesi

Sayın Yrd. Doç. Dr. Salim RAZI
Yabancı Diller Eğitimi Bölümü İngiliz Dili Eğitimi Anabilim Dalı Öğretim Üyesi

Akademik danışmanlığını yürüttüğünüz Okutman Özgür ŞAHAN'a ait "Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi" başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için Fakültemiz Yabancı Diller Eğitimi Bölümü İngiliz Dili Eğitimi Anabilim Dalı'nda görev yapan Arş. Gör. , 01 Mayıs 2016 – 01 Mayıs 2017 tarihleri arasında söz konusu araştırma çalışmasına katılımında Fakültemiz açısından herhangi bir sakınca bulunmamaktadır.

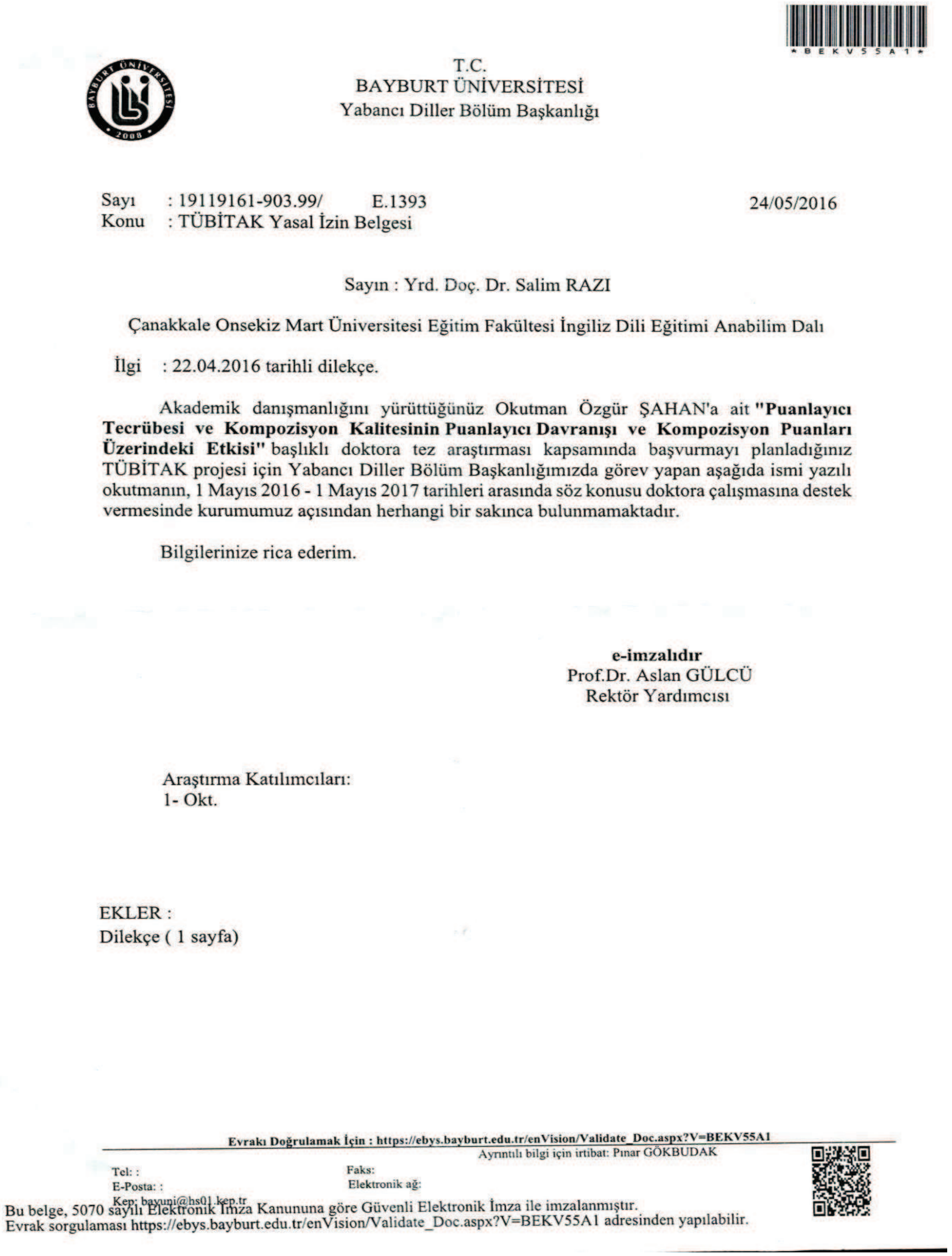
Bilgilerinizi arz/rica ederim.

e-imzalıdır
Prof.Dr. Dinçay KÖKSAL
Dekan V.

Not: 5070 sayılı elektronik imza kanunu gere i bu belge elektronik imza ile imzalanmı tır.

Anafartalar Kampüsü 17100
2862171303

Bilgi için:Alp ARSLAN
Teknisyen



BEKV55A1

T.C.
BAYBURT ÜNİVERSİTESİ
Yabancı Diller Bölüm Başkanlığı

Sayı : 19119161-903.99/ E.1393 24/05/2016
Konu : TÜBİTAK Yasal İzin Belgesi

Sayın : Yrd. Doç. Dr. Salim RAZI

Çanakkale Onsekiz Mart Üniversitesi Eğitim Fakültesi İngiliz Dili Eğitimi Anabilim Dalı

İlgi : 22.04.2016 tarihli dilekçe.

Akademik danışmanlığını yürüttüğünüz Okutman Özgür ŞAHAN'a ait **"Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi"** başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için Yabancı Diller Bölüm Başkanlığımızda görev yapan aşağıda ismi yazılı okutmanın, 1 Mayıs 2016 - 1 Mayıs 2017 tarihleri arasında söz konusu doktora çalışmasına destek vermesinde kurumumuz açısından herhangi bir sakınca bulunmamaktadır.

Bilgilerinize rica ederim.

e-imzalıdır
Prof.Dr. Aslan GÜLCÜ
Rektör Yardımcısı

Araştırma Katılımcıları:
1- Okt.

EKLER :
Dilekçe (1 sayfa)

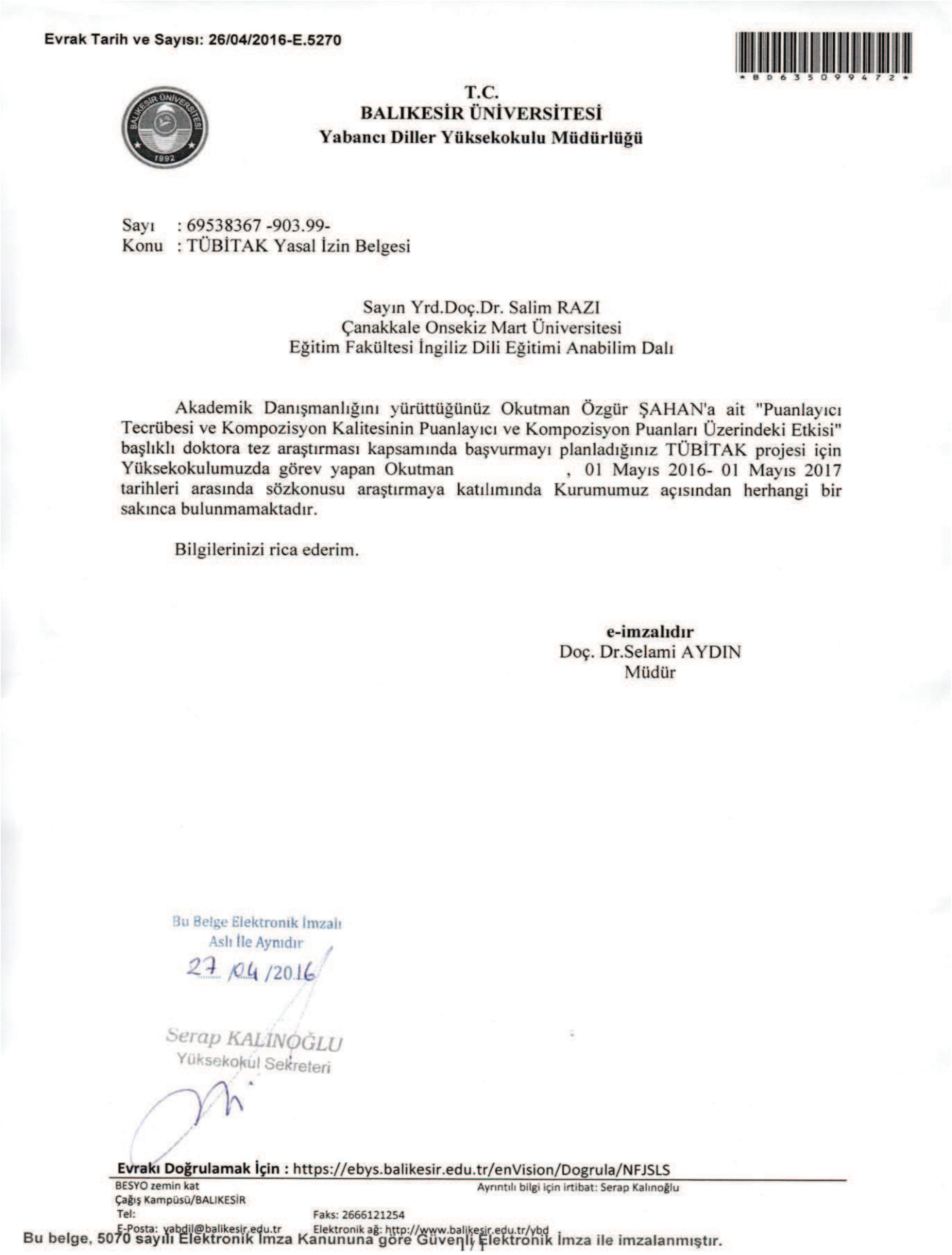
Evrakı Doğrulamak İçin : https://ebys.bayburt.edu.tr/enVision/Validate_Doc.aspx?V=BEKV55A1

Ayrıntılı bilgi için irtibat: Pınar GÖKBUDAK

Tel: :
E-Posta: :
Kep: bayuni@hs01.kep.tr

Faks:
Elektronik ağ:

Bu belge, 5070 sayılı Elektronik İmza Kanununa göre Güvenli Elektronik İmza ile imzalanmıştır.
Evrak sorgulaması https://ebys.bayburt.edu.tr/enVision/Validate_Doc.aspx?V=BEKV55A1 adresinden yapılabilir.



Evrak Tarih ve Sayısı: 26/04/2016-E.5270

BD635099472

T.C.
BALIKESİR ÜNİVERSİTESİ
Yabancı Diller Yüksekokulu Müdürlüğü

Sayı : 69538367 -903.99-
Konu : TÜBİTAK Yasal İzin Belgesi

Sayın Yrd.Doç.Dr. Salim RAZI
Çanakkale Onsekiz Mart Üniversitesi
Eğitim Fakültesi İngiliz Dili Eğitimi Anabilim Dalı

Akademik Danışmanlığını yürüttüğünüz Okutman Özgür ŞAHAN'a ait "Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı ve Kompozisyon Puanları Üzerindeki Etkisi" başlıklı doktora tez araştırması kapsamında başvurmayı planladığınız TÜBİTAK projesi için Yüksekokulumuzda görev yapan Okutman , 01 Mayıs 2016- 01 Mayıs 2017 tarihleri arasında sözkonusu araştırmaya katılımında Kurumumuz açısından herhangi bir sakınca bulunmamaktadır.

Bilgilerinizi rica ederim.

e-imzalıdır
Doç. Dr.Selami AYDIN
Müdür

Bu Belge Elektronik İmzalı
Aslı İle Aynıdır
27/04/2016
Serap KALINOĞLU
Yüksekokul Sekreteri

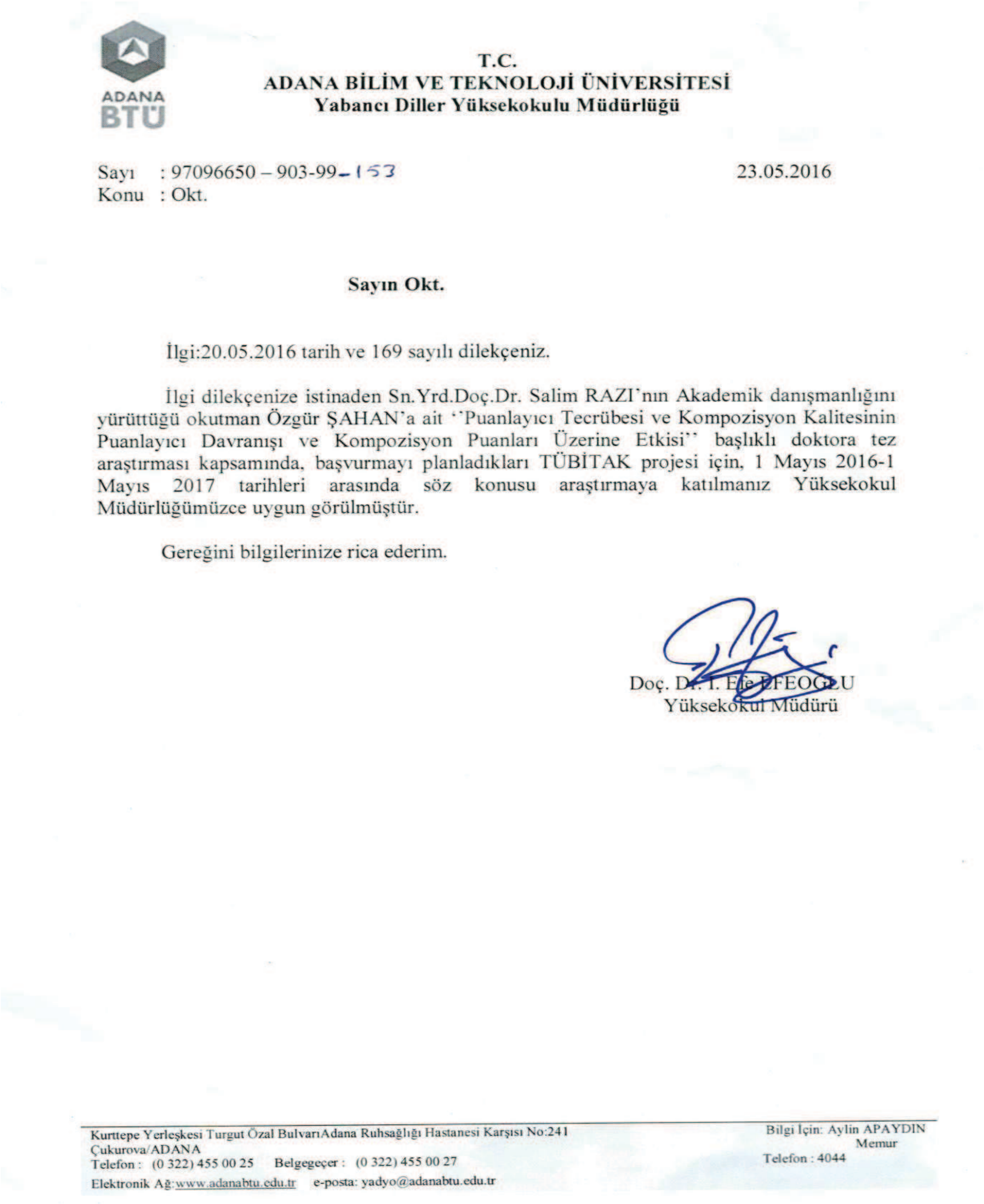
Evrakı Doğrulamak İçin : https://ebys.balikesir.edu.tr/enVision/Dogrula/NFJSLS

BESYO zemin kat
Çağış Kampüsü/BALIKESİR
Tel:
E-Posta: yabdil@balikesir.edu.tr

Ayrıntılı bilgi için irtibat: Serap Kalınoğlu
Faks: 2666121254
Elektronik ağ: http://www.balikesir.edu.tr/ybd

Bu belge, 5070 sayılı Elektronik İmza Kanununa göre Güvenli Elektronik İmza ile imzalanmıştır.

1/1



T.C.
ADANA BİLİM VE TEKNOLOJİ ÜNİVERSİTESİ
Yabancı Diller Yüksekokulu Müdürlüğü

Sayı : 97096650 – 903-99-153 23.05.2016
Konu : Okt.

Sayın Okt.

İlgi:20.05.2016 tarih ve 169 sayılı dilekçeniz.

İlgi dilekçenize istinaden Sn.Yrd.Doç.Dr. Salim RAZI'nın Akademik danışmanlığını yürüttüğü okutman Özgür ŞAHAN'a ait ''Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerine Etkisi'' başlıklı doktora tez araştırması kapsamında, başvurmayı planladıkları TÜBİTAK projesi için, 1 Mayıs 2016-1 Mayıs 2017 tarihleri arasında söz konusu araştırmaya katılmanız Yüksekokul Müdürlüğümüzce uygun görülmüştür.

Gereğini bilgilerinize rica ederim.

Doç. Dr. İ. Efe EFEOĞLU
Yüksekokul Müdürü

Kurttepe Yerleşkesi Turgut Özal BulvarıAdana Ruhsağlığı Hastanesi Karşısı No:241
Çukurova/ADANA
Telefon : (0 322) 455 00 25 Belgegeçer : (0 322) 455 00 27
Elektronik Ağ:www.adanabtu.edu.tr e-posta: yadyo@adanabtu.edu.tr

Bilgi İçin: Aylin APAYDIN
Memur
Telefon : 4044

Appendix M: Descriptive Statistics for Scores Assigned to High-quality Essays

Essay	Range	Minimum Value	Maximum Value	Mean Value	Std. Deviation
1	5.30	4.70	10.00	7.75	1.73
2	5.80	4.00	9.80	7.85	1.55
3	6.50	3.50	10.00	7.23	1.65
4	5.90	4.00	9.90	7.58	1.74
5	7.30	2.70	10.00	7.84	1.68
6	6.00	4.00	10.00	7.67	1.58
7	7.40	2.60	10.00	8.07	1.86
8	6.60	3.20	9.80	7.27	2.02
9	5.90	4.10	10.00	8.53	1.32
10	6.70	3.30	10.00	8.02	1.92
11	6.10	3.90	10.00	7.52	1.70
12	6.60	3.40	10.00	7.30	1.74
13	7.10	2.90	10.00	7.83	1.74
14	4.50	5.50	10.00	7.86	1.39
15	5.20	4.80	10.00	7.80	1.44
16	7.60	2.20	9.80	6.25	1.93
17	6.40	3.60	10.00	8.21	1.68
18	5.50	4.50	10.00	7.75	1.51
19	6.10	3.30	9.40	6.82	1.54
20	8.10	1.90	10.00	7.00	1.75
21	5.20	4.80	10.00	7.95	1.63
22	5.00	5.00	10.00	8.41	1.43
23	4.70	5.30	10.00	7.91	1.24
24	4.90	4.90	9.80	7.46	1.45
25	6.80	3.20	10.00	7.91	1.91

Appendix N: Descriptive Statistics for Scores Assigned to Low-quality Essays

Essay	Range	Minimum Value	Maximum Value	Mean Value	Std. Deviation
26	5.50	2.40	7.90	4.82	1.47
27	6.90	1.40	8.30	4.55	1.82
28	6.80	1.50	8.30	5.28	1.88
29	5.60	2.30	7.90	5.17	1.42
30	6.90	1.90	8.80	4.71	1.50
31	5.90	2.40	8.30	5.01	1.47
32	7.50	2.40	9.90	5.28	1.80
33	7.60	1.10	8.70	4.96	1.68
34	7.40	1.10	8.50	5.18	1.84
35	7.00	2.80	9.80	5.89	1.89
36	5.50	2.40	7.90	4.15	1.38
37	5.40	0.00	5.40	2.68	1.36
38	7.60	0.00	7.60	4.12	1.59
39	6.90	0.80	7.70	4.64	1.79
40	4.80	0.30	5.10	2.72	1.25
41	4.10	0.90	5.00	2.88	1.23
42	7.00	1.60	8.60	5.62	1.48
43	6.30	1.60	7.90	4.52	1.57
44	9.10	0.20	9.30	2.86	1.69
45	6.20	3.00	9.20	6.48	1.79
46	6.80	0.30	7.10	3.46	1.57
47	7.60	0.90	8.50	5.16	1.58
48	7.60	0.60	8.20	3.89	1.65
49	7.00	2.30	9.30	5.83	1.72
50	7.40	2.40	9.80	6.44	1.76

Appendix O: Mean Values Assigned to High-quality Essays by Experience Group

Essay	Low-experienced Raters	Medium-experienced Raters	High-experienced Raters	Total
1	7.14	7.98	8.33	7.75
2	7.63	7.49	8.48	7.85
3	7.04	6.97	7.74	7.23
4	6.98	8.25	7.69	7.58
5	7.59	7.73	8.28	7.84
6	7.64	7.37	8.02	7.67
7	7.65	8.27	8.40	8.07
8	6.98	6.66	8.27	7.27
9	8.45	8.90	8.28	8.53
10	7.52	7.94	8.74	8.02
11	6.72	8.02	8.08	7.52
12	7.08	7.23	7.66	7.30
13	7.70	7.43	8.40	7.83
14	7.68	7.76	8.18	7.86
15	7.95	7.44	7.96	7.80
16	5.95	5.58	7.30	6.25
17	7.94	7.84	8.93	8.21
18	7.50	7.56	8.27	7.75
19	6.31	6.66	7.63	6.82
20	6.58	6.73	7.82	7.00
21	7.95	7.40	8.52	7.95
22	8.31	8.13	8.82	8.41
23	7.69	7.93	8.16	7.91
24	7.29	7.24	7.91	7.46
25	7.63	7.70	8.47	7.91

Appendix P: Mean Values Assigned to Low-quality Essays by Experience Group

Essay	Low-experienced Raters	Medium-experienced Raters	High-experienced Raters	Total
26	4.65	4.99	4.85	4.82
27	4.05	4.23	5.53	4.55
28	4.27	5.56	6.31	5.28
29	5.14	5.08	5.29	5.17
30	4.08	4.61	5.62	4.71
31	4.84	5.15	5.09	5.01
32	5.35	4.93	5.52	5.28
33	4.68	4.72	5.56	4.96
34	4.42	5.20	6.14	5.18
35	5.95	5.50	6.21	5.89
36	3.59	4.59	4.43	4.15
37	1.97	3.08	3.22	2.68
38	3.85	3.80	4.80	4.12
39	4.42	4.28	5.28	4.64
40	2.32	2.69	3.27	2.72
41	2.35	2.82	3.62	2.88
42	5.38	5.31	6.26	5.62
43	4.14	4.63	4.91	4.52
44	2.27	2.97	3.52	2.86
45	5.90	6.68	7.03	6.48
46	3.04	3.49	3.97	3.46
47	4.63	5.17	5.84	5.16
48	3.32	3.95	4.57	3.89
49	5.35	5.21	7.08	5.83
50	6.40	5.79	7.14	6.44

Appendix Q: Extended List of G- and Φ Coefficients for All Raters

All essays (*N* = 50)	**N*Raters***	***Ep*^{2}**	**Φ**
	33	.98	.98
	23	.98	.97
	13	.96	.95
	11	.95	.94
	9	.94	.93
	7	.93	.91
	5	.91	.98
	3	**.85**	**.81**
Low-quality essays (*n* = 25)	**N*Raters***	***Ep*^{2}**	**Φ**
	33	.96	.94
	23	.94	.91
	13	.90	.85
	11	.88	.83
	10	**.87**	**.81**
	9	.86	.80
	5	.77	.69
	3	.66	.57
High-quality essays (*n* = 25)	**N*Raters***	***Ep*^{2}**	**Φ**
	33	.81	.71
	38	.83	.73
	42	.84	.75
	48	.86	.78
	53	.87	.79
	58	**.88**	**.81**
	88	.92	.87
	98	.92	.88
	110	.93	.89
	115	.94	.89
	120	.94	.90

Appendix R: Extended List of G- and Φ Coefficients for Low-experienced Raters

All essays (*N* = 50)	**N***Raters*	***Ep²***	**Φ**
	13	.95	.93
	12	.95	.92
	11	.95	.92
	10	.94	.91
	9	.94	.90
	8	.93	.89
	7	.92	.88
	6	.91	.86
	5	**.89**	**.83**
	4	.87	.80
	3	.83	.75
	2	.76	.67
	1	.62	.50
Low-quality essays (*n* = 25)	**N***Raters*	***Ep²***	**Φ**
	13	.89	.85
	12	.89	.84
	11	.88	.82
	10	**.87**	**.81**
	9	.85	.79
	8	.84	.77
	14	.90	.86
	15	.91	.86
	16	.91	.87
	17	.92	.88
	18	.92	.88
	21	.93	.90
High-quality essays (*n* = 25)	**N***Raters*	***Ep²***	**Φ**
	13	.57	.44
	18	.64	.52
	28	.74	.62
	38	.79	.69
	48	.83	.74
	58	.85	.77
	68	.87	.80
	73	**.88**	**.81**
	88	.90	.84
	118	.92	.88
	148	.94	.90

Appendix S: Extended List of G- and Φ Coefficients for Medium-experienced Raters

All essays (*N* = 50)	$\mathbf{N}_{\mathit{Raters}}$	Ep^2	**Φ**
	10	.93	.92
	9	.92	.91
	8	.91	.90
	7	.90	.88
	6	.89	.87
	5	.87	.84
	4	**.84**	**.81**
	3	.80	.77
	2	.73	.68
	1	.57	.52
Low-quality essays (*n* = 25)	$\mathbf{N}_{\mathit{Raters}}$	Ep^2	**Φ**
	10	.85	.79
	9	.83	.78
	8	.82	.76
	7	.80	.73
	11	**.86**	**.81**
	12	.87	.82
	15	.89	.85
	18	.91	.87
	20	.92	.89
	23	.93	.90
High-quality essays (*n* = 25)	$\mathbf{N}_{\mathit{Raters}}$	Ep^2	**Φ**
	10	.52	.46
	15	.62	.56
	25	.73	.68
	35	.79	.75
	45	.83	.79
	50	**.84**	**.81**
	70	.88	.85
	85	.90	.88
	90	.91	.88
	100	.91	.90
	105	.92	.90

Appendix T: Extended List of G- and Φ Coefficients for High-experienced Raters

All essays (*N* = 50)	$\mathbf{N}_{Raters}$	Ep^2	**Φ**
	10	.95	.93
	9	.94	.92
	8	.93	.92
	7	.92	.90
	6	.91	.89
	5	.90	.87
	4	**.88**	**.84**
	3	.84	.80
	2	.78	.73
	1	.64	.58
Low-quality essays (*n* = 25)	$\mathbf{N}_{Raters}$	Ep^2	**Φ**
	10	.87	.85
	9	.86	.83
	8	**.85**	**.82**
	7	.83	.80
	11	.88	.86
	12	.89	.87
	13	.90	.88
	14	.91	.89
	16	.92	.90
High-quality essays (*n* = 25)	$\mathbf{N}_{Raters}$	Ep^2	**Φ**
	10	.47	.27
	15	.57	.35
	25	.69	.48
	35	.75	.56
	45	.80	.62
	60	.80	.69
	70	.86	.72
	80	.87	.75
	90	.89	.77
	100	.90	.79
	110	.91	.80
	115	**.91**	**.81**
	130	.92	.83
	150	.93	.85
	235	.95	.90

Curriculum Vitae

Personal Information

Name & Surname: Özgür ŞAHAN

Place of Birth: Yerköy/Yozgat

Date of Birth**:** 17 July 1987

Educational Background

Bachelor's Degree: Department of English Language Teaching, Kazım Karabekir Education Faculty, Atatürk University

Master's Degree: Division of English Linguistics, Department of English Language and Literature, Graduate School of Social Sciences, Atatürk University

Doctorate Degree: Department of English Language Teaching, Graduate School of Educational Sciences, Çanakkale Onsekiz Mart University

Publications

a) International Books and Book Chapters

Çoban, M., **Şahan, Ö**., & Şahan, K. E. (2017). A needs analysis based study: The professional development needs of EFL instructors at a technical university. In D. Köksal (Ed.), *Researching ELT: Classroom methodology and beyond* (pp. 229-242). Frankfurt: Peter Lang.

b) International Refereed Research Journals

Şahan, Ö., Çoban, M., & Razı, S. (2016). Students learn English idioms through WhatsApp: Extensive use of smartphones. *Journal of Education Faculty, 18*(2), 1230-1251.

Şahan, Ö., Çoban, M., & Topkaya, E. (2016). A language needs analysis of engineering undergraduate students at a technical university: A Turkish case. *English for Specific Purposes World, 51*(17), 1-33.

Han, T., Tanrıöver, A. S., & **Şahan, Ö. (**2016). EFL students' and teachers' attitudes towards foreign language speaking anxiety: A look at NESTs and Non-NESTs. *International Education Studies, 9*(3), 1-11.

Şahan, Ö., Şahan, K. E., & Razı, S. (2014). Turkish language proficiency and cultural adaptation of American EFL teachers in Turkey. *Procedia Social and Behavioral Sciences, 158*, 304-311.

c) National Refereed Research Journals

Şahan, Ö., & Şahan, K. E., (2014). The relationship between student evaluation of lecturer performance and lecturer self-assessment. *Erzincan Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 7*(2), 85-94.

d) International Conferences:

Han, T., & **Şahan, Ö**. (2017, May). *Training instructors to make reliable judgements of EFL writing.* Paper presented at the 6th International Conference of Strategic Research in Social Science and Education, Prague, Check Republic.

Şahan, Ö., & Razı, S. (2017, June). *The impact of rater experience and essay quality on rater behavior and scoring*. Paper presented at Symposium on Second Language Writing: Assessing Writing, Bangkok, Thailand.

Razı, S., & **Şahan, Ö.** (2017, June). *Integrating peer and self-review skills in the assessment of writing*. Paper presented at Symposium on Second Language Writing: Assessing Writing, Bangkok, Thailand.

Şahan, Ö., & Razı, S. (2017, May). *Investigating rater cognition: How do teachers think while assessing writing?* Paper presented GlobELT Conference, Ephesus, Turkey.

Çoban, M., & **Şahan, Ö**. (2017, April). *An academic writing needs analysis of research assistants at a technical university*. Paper presented at CUELT Cukurova

International ELT Teachers Conferences: Reshaping Teaching and Learning English for the 21st Century, Adana, Turkey.

Şahan, K. E., & **Şahan, Ö.** (2017, April). *What standards for whom? Investigating a preparatory programs' accreditation process.* Paper presented at Towards Higher Education (THE) Conference, Istanbul, Turkey.

Şahan, Ö., & Çoban, M. (2017, March). *Integration of video-enhanced gamification pedagogy into EFL learning: A case study*. Paper presented at STORIES 2017: Students' Ongoing Research in Educational Studies, Oxford, UK.

Şahan, Ö., Çoban, M., & Topkaya, E. (2016, May). *The academic writing needs of research assistants at a technical university*. Paper presented at 9th ELT Research Conference, Interdisciplinary Approaches: Beyond the Borders of ELT Methodology, Çanakkale, Turkey.

Şahan, Ö., Çoban, M., & Şahan, K. E. (2016, May). *A needs analysis based study: The professional development needs of English instructors at a technical university*. Paper presented at 9th ELT Research Conference, Interdisciplinary Approaches: Beyond the Borders of ELT Methodology, Çanakkale, Turkey.

Şahan, Ö., Çoban, M., & Şahan, K. E. (2015, September). *Idioms on your phone: A study on the use of WhatsApp as a language learning tool*. Paper presented at 1st HUMAN Social Interaction and Applied Linguistics Postgraduate Conference, Ankara, Turkey.

Şahan, Ö., & Çoban, M. (2015, June). *A corpus-based study of rhetorical patterns in Turkish university students' argumentative essays*. Paper presented at 18th Warwick International Postgraduate Conference in Applied Linguistics, Warwick, UK.

Şahan, Ö., Çoban, M., & Şahan, K. E (2015, June). *Students learn English idioms through WhatsApp: use of smartphones outside the classroom context.* Paper presented at 18th

Warwick International Postgraduate Conference in Applied Linguistics, Warwick, UK.

Şahan, Ö., Çoban, M., & Topkaya, E. (2015, May). *A language needs analysis of engineering undergraduate students at a technical university: A multidimensional approach*. Paper presented at 3rd ULEAD Congress, International Conference on Applied Linguistics: Current Issues in Applied Linguistics, Çanakkale, Turkey.

Razı, S., **Şahan, Ö.,** & Şahan, K. E. (2014, November). *Turkish language proficiency and cultural adaptation of American EFL teachers in Turkey*. Paper presented at 14th International Association for Languages and Intercultural Communication, Aveiro, Portugal.

Şahan, Ö., Razı, S., & Coffman, K. E. (2014, May). *The impact of Turkish language proficiency on the cultural adaptation process of American EFL teachers in Turkey*. Paper presented at the 8th ELT Research Conference - Innovative approaches to research in ELT, Çanakkale, Turkey.

Şahan, Ö., & Coffman, K. E. (2014, April). *The relationship between student evaluation of lecturer performance and lecturer self-assessment.* Paper presented at 2nd International Congress of Research in Education: Innovative research in education: Implications for future, İzmir, Turkey.

Tanriöver, A. S., & **Şahan, Ö**. (2013, June). *Native and non-native teachers' self- perceptions of and attitudes towards foreign language writing assessment*. Paper presented at 2nd International Symposium on Language and Communication Exploring Novelties, İzmir, Turkey.

Şahan, Ö., & Barın, M. (2012, June). *The impact of direct versus indirect feedbacks to EFL compositions upon the linguistic accuracy and EFL writing motivation.* Paper

presented at 1st International Symposium on Language and Communication Research Trends and Challenges, İzmir, Turkey.

Work Experience

2012 – Present: EFL Instructor, Bursa Technical University, School of Foreign Languages

2009 – 2012: EFL Instructor, Erzincan University, Vocational School of Tourism and Hotel Management

Contact

E-mail Address: ozgur.sahan@btu.edu.tr, ozgursahan66@hotmail.com