THE EFFECT OF STANDARDISATION SESSIONS CONDUCTED BEFORE ENGLISH LANGUAGE WRITING EXAMS ON INTER-RATER AND INTRA-RATER RELIABILITY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MAHMURE NUR KARADENİZLİ-ÇİLİNGİR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF ARTS
IN
THE DEPARTMENT OF ENGLISH LANGUAGE TEACHING

AUGUST 2019

Approval of the Graduate School of Social Sciences

_____

Assoc. Prof. Dr. Saadettin Kirazcı
Director (Acting)

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Arts.

_____

Prof. Dr. Çiğdem Sağın Şimşek
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Arts.

_____

Prof. Dr. Gölge Seferoğlu
Supervisor

Examining Committee Members

Prof. Dr. Gülsev Pakkan	(Selçuk Uni., ELL)	_____
Prof. Dr. Gölge Seferoğlu	(METU, FLE)	_____
Assist. Prof. Dr. Müge Gündüz	(METU, FLE)	_____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Mahmure Nur Karadenizli Çilingir

Signature:

ABSTRACT

THE EFFECT OF STANDARDISATION SESSIONS CONDUCTED BEFORE ENGLISH LANGUAGE WRITING EXAMS ON INTER-RATER AND INTRA-RATER RELIABILITY

Karadenizli-Çilingir Mahmure Nur

M.A., English Language Teaching

Supervisor: Prof. Dr. Gölge Seferoğlu

August 2019, 121 pages

To assess students' writing proficiency, *Indirect* writing assessment, whose evaluation is based on objective judgment of the rater, and *Direct* writing assessment, which requires raters' subjective judgement are conducted. Hence, it is of utmost importance to provide reliability in the grading of the latter one. Therefore, many studies have focused on rater reliability in the related literature. However, there are very few that have studied both intra- and inter-rater reliability at the same time. The current study, hence, differs from most studies in the literature in that it investigates the effect of standardisation sessions on both inter- and intra-rater

reliability. The study was conducted with 24 English Language instructors working in the preparatory school of a foundation university in Ankara. It consisted of two phases. In the first phase, instructors first received a standardisation session and then double-marked writing papers of an actual proficiency exam. After eight months, in the second phase of the study, the same papers were assessed by the same instructors without a standardisation session. The Wilcoxon Signed Ranks Test results revealed that there was a significant difference only between ten instructors' two assessments. However, when the total scores of all the papers (n=240) were studied, a statistically significant difference was found between the pre- and post-test scores. Moreover, when the papers and instructors' grades were examined individually, it was noted that there were serious differences between the scores of the pre- and post-test. Also, it was observed that there were large discrepancies between the grades of the pairing raters in the post-test. Therefore, the study concludes that standardisation sessions are effective in terms of increasing both inter- and intra-rater reliability and it suggests that standardisation sessions be conducted before any evaluation of a qualitative assessment, which requires raters' subjective judgements.

Keywords: inter-rater reliability, intra-rater reliability, standardisation sessions, writing assessment, rater training

ÖZ

İNGİLİZCE YAZMA BECERİSİ SINAVLARI ÖNCESİNDE YAPILAN STANDARDİZASYON TOPLANTILARININ PUANLAYICI VE PUANLAMA GÜVENİRLİĞİNE ETKİSİ

Karadenizli-Çilingir Mahmure Nur

Yüksek Lisans, İngiliz Dili Eğitimi

Tez Yöneticisi: Prof. Dr. Gölge Seferoğlu

Ağustos 2019, 121 sayfa

Öğrencilerin yazma becerilerini ölçmek için, değerlendirmesi puanlayıcının nesnel yargısına dayalı *Dolaylı*; ve öznel yargısını gerektiren *Doğrudan* yazma becerisi sınavları uygulanır. Dolayısıyla, bahsi geçen ikinci ölçme türünde değerlendirme güvenirliğini sağlamak büyük önem taşır. Bu nedenle, literatürde, değerlendirme güvenirliği üzerine birçok çalışma vardır. Ancak, hem puanlayıcı (inter-rater) hem de puanlama (intra-rater) güvenirliğini aynı anda ele alan çok az çalışma bulunmaktadır. Dolayısıyla bu çalışma, standardizasyon toplantılarının, hem puanlayıcı hem de puanlama güvenirliği üzerine etkisini araştırması nedeniyle, literatürdeki birçok

çalışmadan farklıdır. Bu çalışma, Ankara'da bir vakıf üniversitenin hazırlık okulunda çalışmakta olan 24 İngilizce öğretim görevlisi ile yapılmıştır. Çalışma iki aşamadan oluşmaktadır. İlk aşamada, öğretim görevlileri ilk olarak standardizasyon toplantısına katılmış, sonrasındaysa her bir kağıdı iki kez değerlendirdikleri gerçek bir hazırlık atlama sınavının yazılı anlatım bölümüne ait öğrenci kağıtlarını notlamışlardır. Sekiz ay sonra, çalışmanın ikinci aşamasında, aynı öğrenci kağıtları yine aynı öğretim görevlilerince tekrar değerlendirilmiştir. Wilcoxon Signed Test veri çözümlemesi sonuçlarına göre, sadece 10 öğretim görevlisinin iki değerlendirmesi arasında anlamlı bir fark görülmüştür. Ancak, tüm kağıtların (n=240) toplam puanları incelendiğinde ilk ve ikinci değerlendirmelerin arasında istatistiksel olarak anlamlı bir fark olduğu görülmüştür. Ayrıca kağıtlar ve öğretim görevlilerinin notları tek tek incelendiğinde, ön test ve son test sonuçları arasında ciddi farklılıklara rastlanmıştır. Bununla birlikte, ikinci değerlendirmede, aynı öğrenci kağıtlarını değerlendiren partnerlerin notları arasındaki farkın çok fazla olduğu gözlemlenmiştir. Sonuç olarak, bu çalışma, standardizasyon toplantılarının hem puanlayıcı hem de puanlama güvenirliğini arttırmakta etkili olduğu sonucuna ulaşarak, bu toplantıların, puanlayıcıların öznel yargısını gerektiren tüm nitel ölçütlerin değerlendirmesinde uygulanması gerektiğini savunmaktadır.

Anahtar sözcükler: puanlayıcı güvenirliği, puanlama güvenirliği, standardizasyon toplantıları, yazılı anlatım değerlendirmesi, puanlayıcı eğitimi

To my beloved daughter, Deren…

ACKNOWLEDGEMENTS

I would like to offer my sincere gratitude to everyone who helped and encouraged me to complete this journey of mine.

First and foremost, I would like to express my endless gratitude to my dear advisor, Prof. Dr. Gölge SEFEROĞLU, for her invaluable guidance and inspiration. Whenever I needed her, she was always there to support me with her constructive feedback and encouragement. Her continuous guidance throughout this journey made me believe in myself and motivated me to complete this thesis. I am also indebted to her for her patience and understanding attitude towards me. Without her, I would not have been able to complete this study.

I also owe many thanks to my committee members, Prof. Dr. Gülsev PAKKAN and Asst. Prof. Dr. Müge GÜNDÜZ, for their precious time, advice and support.

I wish to extent my gratitude to all my teachers in the Department of English Language Teaching at Middle East Technical University, especially to Prof. Dr. Martina GRAČANIN YÜKSEK, Assoc. Prof. Dr. Bilal KIRKICI, Assoc. Prof. Dr. Perihan SAVAŞ, Assoc. Prof. Dr. Nurdan ÖZBEK GÜRBÜZ, Dr. Deniz ŞALLI ÇOPUR and Dr. Işıl Günseli KAÇAR. I am indebted to each of them since they taught me a lot in their courses and enabled me to successfully complete my journey at Middle East Technical University.

My heartfelt thanks go to my dear colleagues who work at the University where this study was conducted. They kindly accepted to be a part of my study

despite their hectic workload. Without their precious help, it would not be possible for me to conduct this study. I will always be indebted to them for their precious time and effort they put in this study. Also, I would like to thank Joseph Blake LEWIS, John Michael FLYNN and Julia GOGGIN for proofreading and reviewing this thesis. I am also deeply thankful to Sercan ÇELİK for his invaluable support and time. Whenever I needed him, he was always there to help me. *Thank you!*

I would also like to thank all of my friends in my life, but especially my dearest one, Sinem SAATLİ ÖZER, for being a sister to me with her invaluable advice, support and guidance all the time. She has always been the torch to enlighten my path whenever I need her. *Thank you!*

My special gratitude goes to my husband, Arslan ÇİLİNGİR, and his parents, who patiently took care of my little one while I had to study on my thesis. *Thank you!*

Last but not least, I would like to thank my beloved family, Cengiz KARADENİZLİ, Nurselen KARADENİZLİ, Gökberk KARADENİZLİ and Saadet Nur KARADENİZLİ for their endless faith in me and unconditional love for me. I am lucky to have such a great family. *Thank you!*

And my dear daughter Deren! Thank God I have you! You are my miracle and the most precious present given to me in this life. As I always tell you, *thank you for being 'my' daughter!*

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.0 Introduction

This chapter briefly provides the theoretical background for the present study. Then, it presents the significance and the scope of the study respectively. Finally, the research questions are provided.

1.1. Background to the Study

In the field of second/foreign language teaching, assessing writing skill has been considered a challenging task as finding an assessment method which is both reliable and valid is quite difficult. Test reliability and validity are two essential features of a test that denote the quality and efficacy of the test. However, as many scholars in the field suggest, the methods that have been employed to measure writing ability are not always satisfactorily valid or reliable. While reliability refers to the consistency of a measure, validity indicates how well a test measures what it is supposed to measure. In other words, ‘a reliable method of assessing writing ability would yield a consistent judgment of a student’s abilities if applied again […] and a valid writing assessment would be sensitive to a writer’s "true" abilities’ (Charney 1984). For Hughes, a proper writing task should; (1) elicit students’ true writing ability, (2) involve a set of tasks students are supposed to perform, and (3) be

appropriately scored by teachers (2003, p. 83). He, however, suggests that the most problematic part of all three is to develop a scoring procedure.

The current methods of writing assessment are classified as indirect and direct assessments. Indirect methods may involve diction, spelling, grammar, punctuation, syntax, sentence order and some aspects of style (Charney, 1984). In the literature, indirect writing assessments are also referred to as *quantitative methods* as the rater's scoring is not open to interpretation but rather based on an objective scale[1], therefore, they offer high scoring reliability. However, though considered to be reliable measures, they were not eligible enough to measure writing ability by themselves (Cooper, 1977). Therefore, this paved the way for the advancement of direct writing assessment. In a direct writing measure, students are asked to produce a text based on a given topic, therefore, allowing makers to evaluate the high writing skills of students constructed on more valid criteria (Charney, 1984). Essay-based exams are one popular form of this assessment type. Direct assessment types also denote *qualitative methods* in the literature since the rater's subjective interpretation and judgement are required during scoring[2]. For many leading names in assessment like McColly (1970), diverse subjective judgements mean varying protocols employed by raters during scoring. To put simply, different markers will focus on different aspects of a piece of writing which will cause diverse results. Consequently,

[1] So as to keep the citations as they appeared in the original, the term *quantitative method* was interchangeably used for the indirect assessment type in some citations.

[2] So as to keep the citations as they appeared in the original, the term *qualitative method* was interchangeably used for the direct assessment type in some citations.

direct measures may not always produce accurate and reliable results and cannot be trusted all the time.

So what are the factors that affect rater reliability? Rater bias Venkatasamy (2016) and educational background (Venkatasamy 2016 and McNamara 1996) play an important role in jeopardizing rater reliability. Shi (2001) suggests rater language background is an essential factor which influences rater reliability. Her study found that although native and non-native teachers gave similar scores to the students and achieved high reliability in their rating, both groups focused on different areas while grading (Venkatasamy, 2016). Another major factor affecting reliability is the amount of experience the raters have: inexperienced markers tend to mark more inconsistently than experienced markers (Ghanbari, Barati & Moinzadeh, 2012; Venkatasamy, 2016 and Vergeer & Eiting, 1997).

Many studies have been conducted concerning how to eliminate, or reduce, the factors that decrease rater reliability. First, a number of studies agree using a rubric contributes a lot to improving reliability between two markers' assessments, in other words *inter-rater reliability* (Jonsson & Svingby, 2007; Kohn, 2006; Silvestri & Oescher, 2006) as well as *intra-rater reliability*, which refers to the assessment of the same work by the same rater at different times (Moskal & Leyden 2000). Using rubric helps raters revisit predetermined set of criteria whenever they need consultation, which results in better consistency and higher reliability in their judgement.

Second, multiple/double marking protocol was suggested to improve reliability. Starting first in a completely different field from language education, Biology (Head, 1966 and Lucas, 1971), the efficiency of multiple/double marking strategies in assessment was researched in diverse fields. It was Wood and Quinn that investigated its impact on English writing assessments. They stated that it would lead to better reliability and consistency between markers although it would not eliminate inconsistency altogether and that some disagreement between markers would also be tolerated (1976).

Although all these attempts contributed a lot, many in the field believed they could not yield the desired level of rater-reliability *per se* and suggested they be assisted with training in this area. In the literature, these trainings have different names, such as *rater training*, *standardisation sessions*, *calibrating meetings* or *norming sessions*[3] - though they all refer to the same practices. These trainings could be a part of an extensive assessment-related program or they could be conducted before each and every assessment of a productive skill. They are mostly given by testing experts, teacher trainers, testing institutions or Testing Unit members at universities.

Training of teachers for productive skills is believed to reconcile teachers' individual judgements as much as possible by providing them with clear understanding of the rubric and the criteria of the exam that teachers need to take

[3] In the related literature, these sessions are mostly called *rater training*. However, because the university in which the study was conducted calls these sessions *standardisation sessions*, the researcher prefers to use this term rather than *rater training*. However, so as to keep the citations as they appeared in the original, the term *rater training* was also interchangeably used for *standardisation sessions* in many citations, especially in Literature Review Chapter.

into consideration during their assessment. It aims to familiarize raters with the scoring procedures and prepare them for dealing with the unexpected situations during the rating (McNamara, 1996, p. 92). Briefly, it is believed to reduce rater variability deriving from raters' different amount of experience (Shohamy, Gordon & Kraemer, 1992; Weigle, 1999), and L1 backgrounds (Brown, 1995) and it helps examiners familiarize themselves with the rubric and the exam specifications.

1.2. Significance of the Study

As one of the productive skills, writing is of great importance in English Language Teaching. A successful writing assessment requires many steps, such as giving students level-appropriate writing instructions, conducting the exam and, finally, scoring it. Practising the last step is usually more complicated than the first two since rating a qualitative assessment is subjective and scores may change from one person to another, or even the same raters may show variety in their judgments in time. Therefore, literature on assessment discusses how to eliminate or diminish the inconsistencies in grading. It presents a number of methods like using a rubric, double marking procedures and standardisation sessions, which help raters follow similar protocols while grading papers.

Since achieving reliability in writing assessment is one of the major objectives of many language schools and higher education institutions, the current study might be a guide for EFL (English as Foreign Language) and ESL (English as a Second Language) teachers who are interested in effectiveness and efficiency of

standardisation sessions. Furthermore, testing and curriculum units in these institutions that conduct such meetings before every writing assessment may also make use of the findings of the present study. Moreover, although there are a number of research studies on how to obtain reliability in writing assessment, it seems that more studies are needed to investigate to what extent standardisation meetings are effective in adjusting raters' judgements. Last but not least, to the knowledge of the researcher, studies in the related literature have mostly focused on either intra- or inter-rater reliability and there are very few that have studied on both at the same time. The current study, thus, differs from similar research as it explores the effects of standardisation sessions on both intra- and inter-rater reliability at the same time. Therefore, the researcher believes that this study may take a unique place in the related literature to provide more guidance in the field.

1.3. Purpose and the Scope

The present study aims to shed light on the effectiveness of standardisation meetings conducted before written assessments. To this end, 24 English Language instructors at the Preparatory English School of a foundation university in Ankara graded an actual proficiency exam in September. Teachers were first asked to attend a face-to-face standardisation session in which the rubric and the exam specifications were made clear to the teachers. They assessed the papers right after the session. Each instructor graded 22 student papers and all of the papers were double-marked. After an eight-month break to avoid carry-over effect of the previous assessment,

among the 22 papers that each instructor graded in September, 10 papers were randomly selected for the reassessment. This time, papers were assessed without any standardisation session. In this phase, the instructors were asked to double-mark the papers as well. The assessment scores of two different markings were analysed quantitatively. The findings were discussed to find out the effect of standardisation meetings on intra- and inter-rater reliability.

The present study is guided by the following research questions:

1. To what extent do standardisation sessions conducted before writing assessments affect *intra-rater reliability*?

2. To what extent do standardisation sessions conducted before writing assessments affect *inter-rater reliability*?

CHAPTER 2

LITERATURE REVIEW

2.0 Introduction

This chapter starts with defining what written assessment means and explaining the different types of it. Then, it discusses why, as a qualitative assessment type, it is difficult to assess it. Finally, it scans and discusses the literature that suggests the methods to standardise examiners like (1) using a rubric, (2) following multiple/double marking protocols; and (3) conducting rater training or, in other words, standardisation sessions.

2.1. Assessment of Writing

In the modern world in which English is *lingua franca*, it is of great significance to gather evidence of what and how much students have learned the language. Therefore, assessing to what extent learners of English have acquired the language has been one of the central issues in higher education recently. To this end, many procedures to evaluate students' learning have appeared and written assessment is one of them. Some of them are indirect assessment of writing like multiple choice, true/false or matching questions while some others, such as essays

in which students write a response to a prompt, directly measure language abilities. Recently, the latter assessment type is more preferable as they are considered to evaluate all the language traits in one exam at the same time. In other words, indirect assessment usually focuses on students' proficiency on specific domains of language knowledge, such as grammar or vocabulary (Perkins, 1983), while direct assessment types more easily address broader learning outcomes in depth as well as allow students to express themselves in a more authentic context.

Although written assessment has always been applied at schools, assessment *through writing* rather than *assessment of writing* was the main focus until recently (Hamp-Lyons, 2002). The direct assessment of writing gained particularly momentum when Harvard University introduced written composition to its entrance exams in 1870s, rather than applying the traditional oral examination (Lunsford, 1986). For Huot (1990), though, the main shift in direct writing assessment started in the mid-1960s when serious research in direct writing assessment began to appear and its popularity culminated in the mid-1970s.

One difference between indirect and direct assessment in writing is how the questions are presented and how the responses are produced. In other words, the question type and the answers expected from students are different in these two types of assessments. Another difference lies in their scoring. Indirect writing methods are objective assessments whereas direct writing assessments rely on raters' subjective interpretation and judgement. That is why indirect written assessments are high in

terms of scoring reliability while direct methods are criticised for their lack of rater reliability.

2.2. Indirect and Direct Assessment of Writing

Current methods of writing assessment can be categorized as indirect and direct. As in Charney, they are also called ‘quantitative’ and ‘qualitative’ assessments of writing respectively (1984). In indirect writing assessments, students are given a number of choices from which they are supposed to choose or match the best option. Put another way, students are to select the appropriate answer from a number of faulty options. Among such assessment types are multiple choice, true/false or matching questions. They aim to measure the conventions of writing such as grammar, spelling, punctuation, sentence construction and the like (Stiggins, 1982). Namely, they do not demand students’ exact writing skill but test students’ knowledge and recognition of effective writing components.

Indirect writing assessment methods are usually easy to administer. They are practical to evaluate students’ language proficiency in a short time. Raters of such measures do not use their discretion but a pre-prepared answer key to mark examinee’s responses. Such test types can be scored in two ways; via optical readers or the personnel who are provided with an answer key after the exam. Therefore, grading either through a machine or teachers who hand score is not time-consuming. Moreover, thanks to their objective scoring, they are believed to have high reliability,

which is one of the two most important concepts in test assessment - along with validity.

- *Reliability*

Reliability is defined as the consistency and accuracy of the assessment tool to measure students' performance (Bachman & Palmer, 1996). For Hughes, "the more similar the scores are, the more reliable the test is said to be" (2003, p.29). Namely, a reliable writing assessment should be able to consistently measure language ability when it is applied to different students; or the same student on different occasion; and it should yield similar results when graded by different raters (Stiggins, 1982). There are two types of reliability: inter-rater and intra-rater reliability. Inter-rater reliability refers to the consistency between two or more markers' assessments, while intra-rater reliability is used for the consistency between the assessments of the same work by the same rater at different times.

For Henning (1991), the reason for low reliability lies in the scoring procedures. Some important names in the literature on assessment suggest that, to be able to provide high reliability, the assessment type should be free of subjectivity. Thus, teachers and administrators need to adopt such a policy that all students are graded objectively and fairly. Recognized studies in the field by Noyes, Sale, and Stalnaker (1945), Diederich (1950), and Godshalk, Swineford, and Coffman (1966) argued it is only direct assessment methods that could offer high reliability. Similarly, Huddleston discussed that since multiple choice measures afforded better reliability, they were superior to and sounder than essays (1952). Reliability issues

and the related literature will be discussed thoroughly in section *2.3 Reliability Issues in Direct and Indirect Assessment of Writing*.

Direct assessment of writing, on the other hand, requires the examinee's actual writing. It aims to measure all the sub-skills of the language like grammar or vocabulary use, which indirect methods also assess, as well as other components of writing, such as organization, or understanding of purpose. Hamp-Lyons classified written assessment types as assessment *through writing*, referring to indirect measures, and *assessment of writing*, which means direct measures (2002). The first systematic studies on shortcomings of indirect assessments and benefits of direct measures appeared in the 1960s and culminated in the mid-1970s. Since then, there has been a growing consensus that there needs to be a shift from indirect to direct assessment of writing. The most common form of direct assessment type is essay-based tests. Students are given a prompt and asked to respond to it. They are graded based on a rubric, which is usually either holistic or analytic. Because raters use their own judgements and discretions during assessment, scoring of such exams rely on examiners' subjectivity. This is what opponents of direct assessment have mostly fought against. Many leading names in the field like Diederich, Huddleston, Godshalk and Stalnaker disagreed with the shift to direct assessment, arguing that such assessments are not reliable because of their scoring protocols. They claim that direct assessment types put scoring reliability at stake, thus, indirect measures are superior to and more reliable than direct assessment types. However, supporters of direct assessment methods insist that a test of writing should be able to measure the

examinee's actual ability to write, therefore, it should require the examinee to write. To this end, they brought on another important issue: validity in assessment. They proposed that indirect methods undermine validity, which is another important conception in test assessment besides reliability.

- *Validity*

Dictionary definitions of validity are "the state or quality of being valid" (Random House Kernerman Webster's College Dictionary, 2010); "being valid, sound, and defensible; showing no inconsistency or deficiency when put to the test" (Picturesque Expressions: A Thematic Dictionary, 1980); "being free from logical flaws or being based on valid reasoning" (American Heritage Dictionary of the English Language, Fifth Edition, 2016). In terms of assessment, validity refers to the accuracy of an assessment and whether or not it measures what it is supposed to measure. Namely, a test is said to be valid if it measures accurately what it intends to measure (Hughes, 2003, p.22). Many opponents of indirect assessment methods argue validity is the central quality for meaningful and fair writing tests, which means that a writing assessment task must judge what it claims to assess and what has been taught (Luo, 2015). Therefore, it is invalid to give a writing task to students if the task does not ask students to write as this task does not serve its purpose (ibid, 2015). Likewise, Cooper indicates that though considered to be reliable measures, they [indirect assessments] are not recognized as the primary writing assessment because they are not eligible to assess students' writing ability by themselves (1977).

Now that a test is valid as long as it assesses what it aims to measure, opponents of indirect writing assessment argue that a test of writing ability should make students write. Therefore, only can direct writing measures serve this purpose, which is 'the actual assessment of writing competence.' In this way, a writing test is able to measure all the desired outcomes at once, including sub-abilities that indirect methods are also able to. They submit that, for instance, multiple choice questions cannot make learners go beyond just recognizing errors and choosing the best option. They cannot make students express themselves using all language skills at once. Brown argues that multiple choice questions require a passive mental state while essays force learners to be mentally active as they make students use other components of writing like organization or mechanics (1978).

Bachman focused on three types of validity: (1) content validity; (2) construct validity; and (3) criterion validity (1990). *Content validity* refers to the content coverage in a test. It refers to what extent a measure reflects a specific domain of content (Greenberg, 1992). Namely, content validity means whether and how well a test is able to reflect the quality or skill that is related to a said-behaviour. *Construct validity* determines whether and how well a test is able to measure what it is supposed to measure (Bachman, 1990). Therefore, the test content should measure the ability that it is supposed to measure (Bachman & Palmer, 1996). To put it another way, the test should be constructed in a way that it should be able to measure what it claims to measure. Finally, *criterion validity* indicates the meaningful relationship between test scores and other indicative criteria (Bachman, 1990). It

refers to the extent the test score correlates with the score of another measure or to the practice in which two different subject groups are tested with the same measure simultaneously. There are two types of criterion validity: concurrent validity and predictive validity. *Concurrent validity* measures how well a specific test correlates with a previous test which proved to be valid and well-established. *Predictive validity*, on the other hand, refers to the extent to which a test score predicts a future performance. The most common practice in which predictive validity is usually used is university entrance procedures in some countries. Students are placed in a university based on their high school grade averages. Their high school grades are claimed to predict their future performance at university. In other words, students with high grade averages at high school are expected to be successful at university, as well.

In addition to the validity types mentioned above, face validity is also frequently referred to in the field. *Face validity* refers to the extent to which a test is able to measure the said-behaviour or construct that it claims to measure. To put it simply, it means whether or not an assessment is eligible enough to test what it is supposed to test. Opponents of indirect assessments of writing, for instance, claim that indirect measures are lacking face validity as these tests do not ask students to *write*, which is the only way for students to demonstrate their true writing ability.

The main focus towards the end of the 1970s and during the 1980s was on the validity issue in writing assessment. Until then, teachers and administrators had already been complaining that multiple choice tests, or other indirect measures, had

been underestimating writing skills and they would lead learners to separately memorize pieces of the language, such as punctuation, spelling or use of language (Witte, et. al., 1986, cited in Greenberg, 1992). Bachman's arguments especially on validity helped these voices be heard more. First, they insisted indirect assessment methods were lacking content validity as they would not require the process of composing, revising and editing ideas which were some chief components of writing (Brown, 1978 and Cooper, 1977). Furthermore, they noted that these types of exams were lacking construct validity. Namely, indirect tests were inadequate in measuring examinees' writing ability by not asking them to write but forcing them to choose the correct answer among faulty options or matching statements. As Brown mentioned in 1978, learners' writing ability was intended to be measured without requiring them to write even a single word. Now that validity means whether and how well a test is able to measure the said behaviour, opponents of indirect assessment argued that it was impossible for indirect writing measures to evaluate students' writing proficiency without asking them to write. Therefore, indirect tests could not provide validity in this sense.

Proponents of indirect assessments argued indirect writing measures had criterion validity because they could provide high correlation with other types of measurements like course grades. This suggestion was refuted in a way that course grades of learners might be affected by many variables, such as course motivation, attendance or diligence so these scores could not be a pure sign of being a proficient writer of the language (Greenberg, 1992). Moreover, Diederich supported that the

results of essay tests often highly correlated with objective tests scores so indirect measures could also provide high criterion validity. However, Berent, Samar and Kelly concluded that indirect writing measures could not provide high criterion validity for each and every population of language learners (1996). Because indirect methods were a mixture of reading and writing competency, they already lacked face validity for deaf and hard-of-hearing examinees, therefore, it was impossible to establish indirect assessments as a measurement tool for this population.

2.3. Reliability Issues in Direct and Indirect Assessment of Writing

In his article, Dobrić briefly discussed the timeline of how direct writing assessment gained momentum (2018). As a part of the Socratic approach, oral-based examinations had been favoured in higher education in the Western world until the mid-1800s. In 1840, it was schools in Boston that first moved from oral-based assessment to written tests so as to fortify objectivity and a standard assessment policy. After a short period of time, in the USA, Harvard University included an essay writing part in its entrance exam in 1874. It was at this time that criticisms about direct assessment methods were first raised in the country. Among criticisms directed to this type of assessment were its lacking practicality (ibid, 2018) and objectivity as its scoring relied on pure judgement and discretion of the rater (Sheppard 1929 and Starch & Elliot 1912). In the same vein, Traver and Anderson (1935), and Stalnaker (1936) pointed out the unreliability of direct test methods during the 1930s. Another notable study not favouring direct measures belonged to

Huddleston (1952). In his Ph.D. dissertation, he gave 763 high-school students an English exam including sections of objective, verbal and essay-based tests. He stated that reliabilities of the objective test (.78) and the verbal test (.96) were satisfactory while the reader reliability of the essay part was quite lower (.62). As a result, he concluded multiple-choice tests were sounder than essay-based tests as the former had higher rater reliability.

Objections to the use of direct writing assessment, though, culminated in 1961 with the study by Diederich, French and Carlton. Briefly, 300 essays of college freshmen were to be graded by 53 raters grouped in five in terms of their professions: English teachers, social science teachers, natural science teachers, writers and editors, lawyers and businessmen. The researchers found out that, while grading the papers, participant raters emphasized five different factors - or *schools of thought* as the researchers called in the study- which were (1) ideas, (2) form, (3) mechanics, (4) wording and (5) flavour (style). Some raters praised delivery while another group attached importance to language use and some others to mechanics. The inter-correlation among these five factors was only .31. Therefore, considering all the results obtained, the study of 1961 concluded that direct writing assessment could not be used as a reliable method to assess students' language proficiency.

However, although the study disapproved direct assessment because of its lack of inter-rater reliability, there were flaws in the study (Braddock, Llyod-Jones & Schoer, 1963 and Dobrić, 2018). Neither the raters were informed about the criteria

nor was a proper training given to them. They were left on their own to grade the papers within a specific period of time.

The 1970s witnessed a larger number of academic community members complaining about the inadequacy of indirect methods. Suggesting these assessments did not conform to the practices in writing classes, they declared these methods null and void due to their ineligibility to assess writing skill:

> Their [indirect measures'] primary function is to rank order people on a scale. This leaves us again with no absolute knowledge about writing ability and a slight sense of embarrassment when we tell people we'll test their writing ability by not requiring them to write a single word. […] even the very best of them test only reading, proofreading, editing, logic and guessing skills. They cannot distinguish between proofreading errors and process errors, reading problems and scribal stutter, failure to consider audience or lack of interest in materials manufactured by someone else. […] And since capacity to recognize problems in other people's writing does not insure capacity to avoid them in one's own writing-especially first draft writing-we can never be sure what the final scores on such tests mean, let alone the subscores. There are even more insidious aspects to multiple-choice writing tests. They require a passive, reactive mental state when actual writing requires and fosters a sense of human agency, an active state. And they are necessarily incomplete, leading the student and perhaps even the teacher to believe that those aspects of writing most easily tested- sentence structure, word meaning, spelling, punctuation and outlining- are the most important to teach and learn. […] No, an objective test all by itself is not a very good measuring device either; it tells us something, but not enough that is concrete (Brown, 1978, p.1-4).

Therefore, the search for *whether or not* direct assessment was reliable left its place to *what causes* lack of reliability among raters. Coffman (1977), for instance, suggested that the scoring differences might be the result of inconsistencies either between different raters or between judgements of the same rater from one time to another:

> […] Some error [in essay examinations] is the result of differences between raters. Some is due to variability in the judgments of rater from one time to another. Both inter-individual and intra-individual variability can be further broken down into at least three components. The extent to which any of these various sources of error are present depends on how the essays are prepared, on how the responses are rated, and on how the scores are used. Awareness on the part of the teacher of the factors contributing to unreliability, an awareness that can be increased by simple experiments, is a first step in improving the reliability of the teacher's essay examinations (p. 36).

2.4. How to Obtain Higher Reliability in Direct Writing Assessment

Now that writing proficiency means "the ability to discover what one wishes to say and to convey one's message through language, syntax and content that are appropriate for one's audience and purpose" (Odell, 1981, p. 103), and it has been widely accepted that it is only direct assessment tools that can achieve this, obtaining high scoring reliability in these measures has been the main attention of the academic community. To this end, many in the field have been focusing on the possible reasons that could affect rater reliability.

Many studies conclude that raters have certain biases while judging papers, causing reliability problems. Sherwin (1969, cited in Sweedler-Brown, 1985) believed that "while graders may look for the same qualities in a good essay, they tend to favour some over others as major determinants to the overall score" (p. 49). Similarly, Scheafer (2008), in a study with lay raters, and Schoonen, Vergeer and Eiting (1997), in research conducted with both lay and experienced raters, revealed that teachers interpreted the rating scale differently. For instance, some graded content severely while some others assessed language use or organization more harshly. In the same vein, in Eckes (2008), the participant raters fell into six different categories in terms of the importance they attached to the scoring rubric. The results of the aforementioned studies were similar to those in Diederich, French and Carlton (1961), which also categorized the participant raters that favoured one aspect of writing over the others. In another study exploring if rater's L1 background affected rater reliability, Shi (2001) included 46 experienced teachers, half of whom were native speakers of English and Chinese speakers of English. Teachers were asked to use a holistic rubric and comment on their grading. The study found out that although both groups scored the papers similarly, they were different in their comments: non-native speakers attached more importance to organization and content compared to native teachers, and they scored these elements more severely. The study concluded that both parties had a different understanding regarding the aspects that a good piece of writing should bear. Therefore, the findings of the study brought up the issue that language background of teachers might be one factor that

could affect rater judgement. Another possible factor to affect reliability is the experience the raters have. Studies in the field have mostly concluded that the less experienced the raters are the lower rater reliability is achieved. For example, Barkaoui (2007) suggested that novice teachers had more difficulty interpreting the assessment criteria properly compared to experienced teachers, therefore, they jeopardized scoring validity. However, he also cautiously stated that even some experienced teachers showed variability in weighing different aspect of the text, which was also similar to the suggestions of Brown (2003), He, Gou, Chien, Chen and Chang (2013), Lumley (2006), Hamp-Lyons and Davies (2008), and also McNamara (1996) who agreed that even experienced teachers tend to impose bias.

A number of studies in the field have also highlighted the fact that the task type may also affect rater judgement. In other words, the same rater may score different task types differently, which risks intra-rater reliability. For instance, the findings of Quellmalz, Capell and Chou (1982) concluded that the participants scored narrative essays much lower than the argumentative ones. Likewise, Carrell and Connor (1991), and He, Gou, Chien, Chen and Chang (2013) found out descriptive task types received higher scores than argumentative ones. All these studies support Goldshalk, Swineford and Coffman (1966), suggesting that the task type affects raters' judgement.

In addition to those mentioned above, the literature offers some other factors that may affect rater judgement. For example, Charney suggests "the number of separate readings of each writing sample, the number of writing samples evaluated

per student, the size of the rating scale […] and the conditions under which the papers are read" possibly affect rater judgement (1984, p. 70). For Shaw and Weir (2007), the reliability estimates are mainly influenced by the number of the readers, scoring methods and scoring range (p. 12). Being a part of community of practice may also influence raters' judgments as Sanderson suggests (2001). Put it differently, markers may be affected by each other while grading papers. Moreover, Laming (2004) and Sadler (1989) argue raters may not be able to make isolated judgements. Simply put, while grading, raters might be affected by the quality of different student works, comparing papers with each other. Therefore, their judgements become comparative rather than being independent for the paper they are grading (ibid, 2004).

It is obvious that there are a number of factors that may lead to rater inconsistency. Knoch, Read and Randow (2007) summarize some other different rater effects as follows:

> (1) *the severity effect*, where raters consistently rate either too harsh or leniently compared to other raters; (2) *the halo effect*, which occurs when raters rate a candidate's performance on the basis of an overall impression, awarding the same score across a number of different rating scales; (3) *the central tendency effect*, avoiding of extreme ratings or preponderance of ratings at or near the scale midpoint; (4) *inconsistency*, which is a tendency of a rater to apply one or more rating scale categories in a way that is inconsistent with the way in which other raters apply the same scale; and (5) *the bias effect*, which is exhibited when raters tend to rate unusually harshly or leniently concerning the one aspect of the rating situation, i.e. consistently rating one category too harshly or leniently (p. 27).

In short, literature agrees that reliable assessment of direct methods is hard work because it is not value-free (Hamp-Lyons, 2002, p. 182) as there are a great number of factors affecting raters 'judgements, such as raters' background (Gonzalez & Roux, 2013; Lim, 2011; Shi, 2001; Shi, Wan & Wen, 2003), range of experience or bias (Black, 1998) and the type of the task (Shavelson, Gao & Baxter, 1996) as well as some other minor factors, such as handwriting of the examinee or the conditions under which the papers are marked. To this end, to be able to eliminate, or reduce, raters' inconsistency, in the first half of the 20th century, many attempts were made to improve the reliability of essay marking (Brooks, 2004).

- *Using a rubric*

Using a rubric is one of these attempts to improve rater reliability. A rubric is a scoring tool for qualitative rating of authentic or complex student work and it includes criteria for rating important dimensions of performance, as well as standards of attainment for those criteria (Jonsson & Svingby, 2007, p. 131).

There are different rubric types and three of them are commonly used. One type of rubric, *holistic evaluation*, is used to assess the overall quality of learners' work and it is usually concerned with the total performance (Rezai & Lovorn, 2010). It is also time-efficient as well as reliable:

> Where there is commitment and time to do the work required to achieve reliability of judgment, holistic evaluation of writing remains the most valid and direct means of rank-ordering students by writing ability. Spending no more than two minutes on each paper, raters can achieve a scoring reliability as high

> as .90 for individual writers. The scores provide a reliable rank-ordering of writers, an ordering which can then be used to make decisions about placement, special instruction, graduation, or grading (Cooper, 1977, p. 4).

Among other well-known rubric types are *analytic scoring*, through which students' writings are evaluated based on different bands, such as content, language use, grammar or mechanics (Diederich, 1974); and *primary-trait scoring* developed by Richard Lloyd-Jones, in which 'scores are based on one or more specific aspects of performance that are essential for the successful completion of the tested task' (Davis, 2018, p.1296).

There are several advantages of using a rubric. First, using a rubric contributes to higher reliability in assessment (Jonsson & Svingby, 2007; Kohn, 2006 and Silvestri & Oescher, 2006). In this way, it increases valid judgement of the performance assessment (ibid, 2007). Similarly, according to Moskal and Leyden (2000), a well-designed rubric is able to enhance consistency between two markers- that is inter-rater reliability- as well as intra-rater reliability, which refers to the assessment of the same work by the same rater at different times. In both cases, the researchers argue, there are possible factors affecting raters' judgements, therefore, using a rubric will help raters revisit a predetermined set of criteria whenever they need consultation. This will render better consistency in their judgement, and thus, higher inter- and intra-rater reliability in assessment. These suggestions are also in line with the findings of Anadol and Doğan (2018). The researchers carried out a study with three separate groups of teachers divided in accordance with their

experience in using scoring rubrics: teachers having less than one year's experience; the ones with more than five years of experience and a mixture of those having little and ample experience in using a scoring rubric. The findings showed that when a well-designed rubric was used, it could yield consistent results regardless of teachers' experience in using rubrics. The study concluded that a well-designed rubric could be trusted in terms of providing inter-rater reliability, so there was no need to form groups of raters who were experienced equally in rating papers. Therefore, using a rubric helps reduce subjectivity (Moskal & Leyden, 2000), and provides raters with uniformity and confidence in their judgements (Spandel, 2006) since they are able to base their judgements on well-defined sets of criteria. Taking into consideration all those mentioned above, it would not be surprising to note that many in the field believe in the positive role of rubrics regarding its enhancement of reliability, consistency and objectivity, concluding that it is better to use a rubric than not to use it (Özel & Acar, 2014; Parlak & Doğan, 2014 and Spandel, 2006).

On the other hand, increasing numbers of people in the field have recently started to stand against the argument that a rubric itself could provide better reliability (Chapman & Inman, 2009; Dawson 2009; Kohn, 2006 and Reddy & Andrade, 2010). A recent example of this argument is the study by Gonzales, Trejo and Roux (2017) in which eleven Mexican EFL teachers were asked to rate papers using an analytical rubric. The findings were in line with a previous study by Gonzales and Roux (2013) and the others who suggested scoring a rubric does not provide reliability *per se*. A similar study yielded the same results (Rezai & Lovorn,

2010). The researchers provided the raters with two different essay samples. The first sample was well-written with correct use of language, spelling and punctuation, however, it could not fully answer the question. On the other hand, the second sample could satisfactorily answer the question while it had many grammatical, spelling and punctuation mistakes. The participants rated both samples once without and once with a rubric. Most raters in the study were affected by mechanics and language use mistakes and even penalized the second sample essay from the content band although it could fully answer the question. The researchers, thus, concluded that using a rubric could not eliminate or reduce bias and subjectivity. These findings were in line with Brooks who agreed that all of these initiatives could not provide the desired results (2004). However, he called Wiseman's report on multiple marking in 1949 a breakthrough regarding the advancement of a new method to overcome inconsistencies among raters (ibid, p. 34).

- *Double and Multiple Marking*

Double marking refers to two examiners' assessing each script independently. The final mark is usually a combination of two separate marks (Meadows & Billington, 2005). Multiple marking, similarly, refers to two or more raters' assessing each script and the final mark's being some combination of the separate marks (ibid).

The first study of multiple marking was conducted by Wiseman with 11-plus candidates in 1949. He combined teams of four people to assess each script of the candidates. The raters did not have to agree with each other and the final score was a combination of those separate four marks. He reported that multiple marking created

high reliability coefficients -even up to 0.95- that could only be obtained in objective indirect assessment types. Therefore, he was the first to claim that multiple marking provides much higher reliability than single marking.

Many other studies followed and shared Wiseman's suggestions in diverse fields, such as Biology, General Studies, Psychology and English Language Teaching (Head, 1966; Lucas, 1971; Maughan & Burdett, 2013; Pilliner, 1969 and Wood & Quin, 1976), suggesting that multiple marking increased reliability. The study of Lucas in 1971, though, suggested a surprising finding that the more the number of raters increased, the less reliability was obtained. Six examiners rated the same 44 Biology scripts and he calculated inter-rater reliability according to how one, two, three and four separate scorings contributed to the final score. In this way, comparative gains from single to double to multiple scoring were assessed. He suggested that multiple marking provides much higher reliability than single marking, however, the greatest leap in reliability was gained through the increase from single to double marker. He argued that as the number of markers increased, the improvement in reliability decreased. This distinctive finding, together with the objections that multiple marking was not cost- and time-efficient, led researchers to study the efficiency of double marking. For instance, Wood and Quin (1976) awarded the same 100 scripts belonging to O-level English language students to 10 markers. The study assured that double marking could also yield reliable results in assessment.

Another study by Chaplen checked self-consistency of raters (1969). Each English essay was marked blindly by two different raters. After three months, they assessed the papers again. The overall reliability of the essays rated at different times was 0.92. The study suggested that double marking be applied as it improves reliability. These findings were similar to Fearnley (2005), suggesting that double marking is effective in providing consistency among raters. More importantly, he also checked whether random pairing and non-blind marking would affect raters' judgements. He concluded that random pairing would increase reliability, however, non-blind marking would decrease it since the annotations of the first marker affected the second marker.

Although there is a great deal of research supporting multiple and double marking to increase reliability, some also question the feasibility and efficiency of this method. The most important problems were logistical problems, as Fearnley put forward (2005). Transporting the scripts from one to another requires additional work and money. Similarly, it was not cost-efficient enough for some exams whose examinee numbers were more than multiple thousands for one session. Meadow and Billington emphasized the difficulty for some awarding bodies to recruit enough examiners to mark papers even for once, let alone twice (2005, p.58). The report presented for Ofqual in 2014 discussed that even if the institutions were able to double the numbers of examiners, it would be inevitable to include huge number of inexperienced raters, which will probably risk the quality of marking. Moreover, whether it was worth spending that much time and money on double marking was

still debatable for many. For instance, Black (cited in Ofqual, 2014) conducted a study under authentic marking conditions with 512,224 marks from 21.562 single marking events. This distinctive study found that, compared to single marking, double marking did not have such a significant effect on improvement on marking. Like Black, many others submitted that double marking might not yield the desired gains in terms of marking reliability and assessment improvement (Bramley, 2010 cited in Ofqual Report, 2014; Brooks, 2004; Fearnley, 2005 and Meadow & Billington; 2005).

In a nutshell, so as to improve reliability, methods like using a rubric and double/multiple marking might not work alone. The Ofqual report in 2014 highlights the change in the literature: the years between the 1940s and 1980s were huge supporters of double marking while the recent years suggest its minimal benefit in writing assessment. Equally, many recent studies mainly argue that rubrics do not improve reliability and consistency *per se*. Unless the rubric, no matter how well-designed it is, and the criteria are negotiated and internalized by the raters, who will either single or double mark, the quality of assessment is still under risk. To this end, many in the field highlight the importance of rater training.

They all assert that, to be able to benefit from rubric maximally, the rubric should be assisted with training sessions (Dempsey, PytlikZillig & Burning, 2009; Gonzales & Roux 2013; Gonzales Trejo & Roux 2017; Knoch, Read & Randow, 2007; Lovorn & Rezai, 2011; Moskal & Leyden, 2000 and Rezai & Lovorn, 2010;) especially for the novice raters (Hitt & Helms, 2009 and Maxwell, 2010).

2.5. The Relation between Rater Training and Rater Reliability

Rater training refers to the sessions conducted to the assessors in an academic field to be able to reduce variability among raters and, thus, improve rater-reliability. In the literature, rater training could be used either as an umbrella term covering any rating-related trainings or could be used interchangeably for standardisation sessions, norming meetings or calibration sessions conducted before every assessment. These trainings could be given by testing experts, teacher trainers or testing institutions. They might also be provided by Testing Unit members at higher education institutions. They could be delivered to teachers either once or periodically as a part of an extensive assessment training program, or iteratively before each marking session of an exam for productive skills. In short, training sessions are conducted before speaking or writing examinations in which marker judgements may vary and jeopardize reliability in assessment.

Training of teachers for productive skills is believed to reconcile teachers' judgements as much as possible by providing them with clear understanding of the rubric and the criteria that teachers need to take into consideration during their assessment. It aims to familiarize raters with the scoring procedures and prepare them for dealing with the unexpected situations during rating (McNamara, 1996, p. 92). During the sessions, the rubric which is used for the specific task and the criteria expected from examinees are explained to the raters by the expert that leads the session. Later, teachers are supplied with exemplar items to define performance categories. In other words, they are provided with quality, average and poor

examinee performances. They are asked to rate them either individually or sometimes in groups, using the rubric delivered to them. At the final step, rater teachers come together to discuss their judgement and its reasons for each performance. They are expected to meet at a common ground as much as possible. It is usually made clear to the raters that some variation is acceptable but raters who consistently rate too high or too low should adjust their standards (Knoch, 2009, p. 31). As a result, the raters are expected to have constituted clearer judgements for similar performances during their forthcoming assessments. In short, standardisation sessions ensure that (1) each assessor consistently makes valid decisions; (2) all assessors make the same decision on the same evidence base and (3) all candidates are assessed fairly (Greatorex and Shannon 2003, p. 3).

To this end, a number of researchers have been in search of effectiveness of rater training sessions in terms of improving rater reliability. Therefore, the search for whether those who have received training can rate more reliably than those who have not is important in the field.

A great deal of research has concluded that rater training is an effective way of reducing rater variability. As aforementioned in the current study, rater background is one of the factors that cause rater variability. Shohamy, Gordon and Kraemer in 1992 had four different groups of raters. In each group, there were 5 raters, all of whom were native speakers of English. Raters differed in terms of (1) their professional background- experienced vs. lay- and (2) their training- one group received training and one did not. Shortly, the four groups of raters were (1)

experienced raters having received training; (2) experienced ones not having received training; (3) lay raters having received training and (4) lay raters not having received training. During the training session held for five experienced and five lay raters, the rubric was reviewed and explained. Different written samples were presented to the raters and they were asked to rate the papers using either holistic scale, analytic scale or primary-trait scale. Raters were asked to negotiate and come to a consensus for their judgements. The study showed that although inter-rater reliability coefficients were quite high, regardless of raters' professional background and training, intra-rater reliability and overall reliability coefficients were much higher for the trained group. The study concluded that rater training had a positive effect on improving rater reliability, not professional background. Therefore, decision makers should be less concerned about raters' background but put more emphasis into intensive training sessions to prepare raters for their tasks (ibid, p. 31). Weigle (1999) reached a similar conclusion. Participant inexperienced teachers rated more harshly than experienced ones before the rater training session. However, after the training session, inexperienced teachers assessed examinees' papers less severely and the difference in severity in both groups decreased to a great extent. Similarly, in 1994, Weigle used verbal protocol for four inexperienced teachers to investigate how effective rater training is. She reported that rater training session was effective in terms of 'bringing participant teachers' judgements in line with the rest in terms of both marks and the procedures by which they arrived at those marks' (cited in Meadows & Billington, p. 51).

Many studies have underlined the fact that non-native speakers rate more severely than their native counterparts. To this end, Brown (1995) had native and non-native speakers of raters to assess speaking skills of examinees. He concluded that, after the training session, judgements of non-native speakers were not as harsh as they used to be and their assessments were quite similar to native speakers'.

Furthermore, many in the field have looked into the effectiveness of using a rubric and concluded that although a well-designed rubric could increase rater-reliability, it must be assisted with rater training sessions as assessors' familiarization with the rubric is significant. For instance, in Gonzales, Trejo and Roux (2017), eleven Mexican EFL university teachers were given five writing samples to rate with an analytical scoring rubric. The researchers noted that using a rubric as well as having a similar background-having the same L1 and working in the same institution- were not enough to improve rater-reliability because raters' judgements showed great variety. Gonzales et. al argued that nine of the participants had previously received rater training and it might have been the reason of the variability among teachers. They finally concluded that a rubric is just a tool to facilitate raters' assessment and only through assessment training or assessment literacy could rater-reliability and validity of students' assessment could be increased (p. 99). These findings are in line with Kayapınar (2014) in which ten experienced ELT teachers assessed essays of EFL students in six different rating sessions, using general impression marking (GIM), essay criteria checklist (ECC) and essay assessment scale (ESAS). Teachers were awarded 44 essays in one batch and 264 essays in total;

and each marking session was held after a 10-week break so as to remove the carry-over effect of the previous assessment. The study concluded that using a rubric did not guarantee inter- and intra-rater reliability *per se* and that 'deliberate training and agreement of raters before any process seems needed' (p. 127). Lovorn and Rezai (2011) conducted a two-phase study with pre-service and in-service teachers. In the first phase, raters were not given any training while, in the second phase participants rated papers after a training session. The researchers suggested that a quality, intensive training in terms of how to use a rubric is required for a more reliable assessment. Another study by the same researchers suggested that rater training be given to raters prior to marking sessions since using a rubric did not work to improve rater reliability *per se* (Rezai, & Lovorn, 2010). Moreover, to be able to investigate the effectiveness of rater training for raters to understand the criteria, the evaluation items and the evaluation procedures, Kondo conducted two speaking assessment sessions with the raters, one with and one without a rater training session (2010). The results indicate that the variance regarding the items was reduced to about one-sixth after training, concluding that through a rater training session, assessors were able to understand the contents of the evaluation much better.

Tajeddin, Alemi and Pashmforroosh (2011) argued that rater training is effective in terms of changing the assessors' attitudes towards the speaking criteria they attached importance before training. In the research study, 28 non-native EFL teachers rated 10 monologs both before and after a rater training program. The rubric they used included different linguistic criteria such as fluency, grammatical accuracy

or vocabulary; and more pragmatic criteria like topic management, organization or intelligibility. The study found out that, in the first rating session, teachers attached great importance to linguistic features in the rubric whereas, in the second phase after training, they redirected their attention more to non-linguistic criteria and assigned more significance to fluency, comprehension and organization. The researchers reported that 'the training program in the study was effective in encouraging teachers to attend to macro-level, high-order components when making global judgement of speaking performance, therefore, the observed changes after the rating training must be embedded in further teacher education' (p. 148).

Fahim and Bijani (2011) explored how judgements of raters are biased towards certain criteria before and after the training program in assessing L2 essays. To this end, 12 EFL raters scored 40 pre-benchmark essays rated by an IELTS trainer. All the essays were typed before they were submitted to the assessors so that raters would not be affected by examinees' handwritings. The differences between the raters' assessments revealed that all the raters became consistent after the training session, which indicated that rater training sessions were effective in reducing, though not completely eliminating, raters' biasedness and severity. These results were also in line with Wigglesworth's previous study which concluded that feedback provided to raters as a part of rater training helped reduce rater biasedness and improve rater-reliability (1993).

In the literature, there are many more studies documenting that rater training sessions are required to reduce rater variability and increase rater consistency and

reliability to a great extent (Elder, Barkhuizen, Knoch, & Randow, 2007; Lumley & McNamara, 1995; Lunz, Wright & Linacre, 1990; Robinson, 2000; Sweedler-Brown, 1985; Wang, 2010 and Weigle, 1998). However, McNamara highlighted the fact that eliminating variability among raters' assessments is not the main purpose of rater trainings. Instead, they help raters to be internally consistent in their own evaluations (1996), which would, as a result, lead to higher inter-rater reliability. Thus, most studies underline the importance of recurring rater training sessions before each and every direct writing assessment. They argue that the effect of a rater training conducted for a specific exam may not endure long enough for raters to be standardised for another one. To this end, conducting a standardising session before each assessment in which exam specifications, criteria and the rubric are discussed is a must to help raters re-establish their judgements. Therefore, on-going standardisation sessions should be a part of teacher training programs at higher education institutions (Lumley & McNamara, 1995).

On the other hand, there are also many others that argue rater training is not that effective. Pufpaff, Clarke and Jones investigated the impact of rater training on the rubric-based of scoring of three pre-service teacher candidates' performance assessment (2015). Three different types of assessments of freshman students - digital portfolios, research papers and case studies- were rated by EFL teachers. They first assessed these tasks with a rubric but in the absence of a rater training session. Later, assessors were provided with an audio embedded within a PowerPoint presentation explaining the re-designed expanded rubrics for each assessment type.

The researchers concluded that rater training had little to no improvement on rater agreement and consistency. However, they cautiously highlighted the limitation of the study that raters completed training on their own, without any supervision of the researchers. They stated that, therefore, they had no idea whether or how long raters spent time with revising the training materials. Similarly, Baird, Greatorex and Bell (2004) conducted a study with 45 raters to explore the effect of standardisation meetings before assessments. They found out that these meetings are not that effective in improving assessors' understanding of mark schemes and improving rater reliability. Moreover, the same researchers stated in their 2002 study that the standardisation session conducted to the participant raters did not help improve rater-reliability although the participants expressed their satisfaction with receiving the session and that they benefitted a lot from it. A previous study by Black (1962) with 19 examiners who marked the same script with a ten-day interval argued that briefing was not effective in standardising the examiners' marking (cited in Meadows & Billington, 2005, p. 51). Similarly, Shaw (2002) examined whether iterative standardisation sessions could help improve rater-reliability. The training included five different sessions, and after each session, raters sent the marked papers back to the researcher. He argued that although inter-rater reliabilities in each batch were high among the assessors (0.77), they did not improve with time and standardisation but remained constant. He highlighted the fact that 'even before the training, raters did not differ grossly from the standard'.

CHAPTER 3

METHODOLOGY

3.0 Introduction

This chapter presents the setting and the participants in the study. Then, it focuses on data collection instruments and procedures. Next, it gives information about the adopted data analysis methods. Finally, ethical considerations are explained briefly.

3.1. Setting

This study was conducted with English Language instructors at the preparatory school of a foundation university in Ankara during the 2018-2019 academic year.

As the medium of instruction at this university is English only, the English Language School (ELS), or the preparatory school, at the university attaches great importance to teaching English in order to allow students to study in their departments successfully. To this end, the ELS administration follows a curriculum that guides its students to toward mastery in all the four language skills, namely listening, speaking, reading and writing.

The levels are determined through a language exam consisting of two stages: English Placement Exam (EPL) and English Proficiency Exam (EPE). The first stage of the exam, EPL, is only given to newly registered students in September before the fall semester starts. The exam questions are ordered from easy to difficult and aim to assess language use and vocabulary knowledge of students. Students who display high levels of language knowledge in the EPL are able to sit the EPE, which is the second stage of the exam. Those who are not eligible to take the EPE are required to start as *Elementary* level students in ELS.

The EPE seeks to assess higher level language skills, such as reading, writing, listening and speaking. Therefore, students sit three different exams. The first part of the exam includes a Reading and Listening section. The second part involves (1) independent writing in which students are required to write on a given prompt and (2) an integrated writing exam in which students first listen to a lecture and then write on a given prompt based on the lecture they have just listened to. The final section assesses students' speaking skills. Each section totals up to 25 and students who are able to score 75 are exempted from the English Preparatory Program and can start in their departments. Those who score under 75 are placed into either Intermediate or Upper level classes based on the total score they obtain from each section in the exam.

The EPE is given to not only new but also previously registered students, no matter their levels in the previous term. To clarify, in most schools, only failing upper-level students in the previous term are allowed to take the proficiency exam in

September. However, at this university, the EPE in September is also open to all previously registered students at any levels. Even if the previously enrolled students failed at the Elementary or Intermediate level in the previous spring semester, they are allowed to sit the EPE. The rationale behind this is to reward students who have spent time studying English while away from school and who have proven they are proficient enough to successfully use all four language skills in their departments.

Furthermore, during each semester at the ELS, no matter at which level (Elementary, Intermediate or Upper), students are expected to take a series of exams that assess their language and vocabulary knowledge, such as quizzes, and exams which assess all four language skills. Students collect points to move to a higher level – or into their department if they are Upper level students. Elementary and Intermediate level students are required to collect 70 points in order to study at a higher level while Upper level students should collect 75 points to begin studying in their departments. If Upper students fail, they are required to take EPE exams which take place in January and May if they expect to start in their departments. The exit level of language proficiency is B2.

In short, it is of utmost importance for the university that all students achieve the determined level of English so that they can successfully continue their studies. The researcher has given such detailed background information to show the current study was conducted in a setting in which English and, therefore its assessment, play a crucial role.

3.1.1 Assessment Procedures

As aforementioned, the EPE involves different sections. The Reading and Listening parts of the exam consist of multiple choice questions. Students use an optical form for their answers and forms are graded by machine. Namely, the assessment of this part of the exam is based on quantitative methods. However, students are also supposed to sit speaking and writing exams with evaluations that demand qualitative methods. To be precise, the assessment of the first part of the exams is objective while the latter group is subjective in nature. Therefore, before grading the writing and speaking parts of the exam, instructors meet in a classroom to reconcile their individual judgements.

The Writing Rubric: At this stage, it is significant to briefly explain the writing criteria used by the ELS. The writing exams are marked based on an analytic rubric developed by the Testing Unit members and then approved by the Testing Head and the ELS administration, respectively. It includes four different bands, which are (1) *Content and Organization* out of 10; (2), *Coherence* out of 5; (3) *Lexical* out of 5, and (4) *Grammar* out 5 points. The minimum score that students can obtain for each band is 1. Therefore, the lowest grade a student may earn from the exam is 5 and the highest is 25. If a student fails to answer the question implicitly or explicitly, or in other words if the paper is off-topic, the paper merits a zero. The rater ticks a band that reads 'The text is unrelated to the topic. The student fails to answer the question implicitly or explicitly.'

In the *Content and Organisation* band, students are expected to answer appropriately the writing question by using convincing supporting ideas without ambiguity and/or overlapping information. In the *Coherence* band, students are required to write a well-organized essay with a clear progression of ideas and appropriate transitions. Their essays are expected to contain an introductory paragraph with a proper thesis statement; one or more body paragraphs that begin with an appropriate topic sentence, and a concluding paragraph. In the *Lexical* and *Grammar* bands, students are assessed in terms of not only accuracy but also the variety of the structures and vocabulary items used.

Standardisation Sessions: The Testing Head or a teacher from the Testing Unit conducts the session. First, they explain and discuss the related writing criteria with other instructors along with exam specifications and the expected qualities for the task in the exam. Later, the teachers are asked to blindly mark the papers that the Testing Unit has chosen beforehand. Approximately five papers are marked blindly by the instructors. Papers are discussed for each and every band in the rubric by the Testing Unit member and the instructors. The ultimate score for each paper is then calculated. Once the standardisation session ends, instructors begin marking the papers. The aim of these sessions is to make teachers meet on a common ground before they start a qualitative assessment, which is subjective. In other words, the examiners need to be thinking along standardised lines so that students may not be advantaged or disadvantaged according to which examiner marks their work (Raikes, Fidler & Gill, 2009).

Double Marking: In addition to standardisation sessions to obtain a high quality marking session, a double marking procedure is followed because the ELS supports that multiple judgments lead to a *truer* final score than any single judgment (Hamp-Lyons, 1990). Once teachers finish scoring a single group of students, they continue to mark papers from a different group, without seeing the first markers' grades. In this current study, the student papers were also double-marked both in the first and the second assessment. Therefore, 24 participant instructors formed 12 pairs. Pairs were not formed by the researcher but by the Head of Testing Unit randomly because the pre-test that was taken in September was the actual proficiency exam. Hence, to be able to see the inter-rater consistency, the researcher kept the pairs in the post-test in May as they were in the pre-test. Each consecutive instructor formed one pair, i.e. Instructor #1 and Instructor #2 were Pair #1 while Instructor #3 and Instructor #4 were Pair #2.

Now that the EPE writing exam is scored on a scale up to 25, the allowed discrepancy between two markers is a maximum of 4. The cases of extreme discrepancies, such as 5 and more, require a third eye for an ultimate assessment. Therefore, to be able to re-evaluate such papers, a Discrepancy Committee, whose members are Testing Unit members, has been established.

Student Papers: 120 students' proficiency exam writing papers were included in the study. Each 24 participants were awarded 10 papers. All the papers were double-marked by the instructors. Therefore, in total, 240 student papers were evaluated twice in the study (24x10=240).

3.2. Participants

The study was conducted with English language instructors at the preparatory school of a foundation university in Ankara. The study involved instructors who were both native speakers (*n*=4) and non-native speakers of English (*n*=20). In total, 24 English Language instructors voluntarily took part in the study, 13 of whom were female and 11 of whom were male. Each of them signed a consent form (Appendix A) and they all filled in a demographic questionnaire that elicited their general background information, teaching and rating experience and, finally, rater training sessions if they had taken any (Appendix B). All non-native teachers held BA degrees in English language related departments. One native teacher held a BA degree in English Linguistics whereas the others held BA degrees in different subjects but had either TEFL or CELTA certifications. Most of the non-native instructors (*n*=18) were still studying for or had already completed their master's degree. Only two of the MA degrees were not English language related; one in Women Studies and one in Advertising.

Table 3.1

Profile Summary of Participant Instructors

Variables	**Categories**	**F**	**%**
Nationality	Native	4	16.67
	Non-Native	20	83.33
	TOTAL	**24**	**100**

Table 3.1 (continued)

Gender	Male	13	54,17
	Female	11	45,83
	TOTAL	**24**	**100**
Degree in BA	ELT	7	29,19
	LIT	9	37,50
	LING	5	20,83
	TRANS	1	4,17
	Others	2	8,33
	TOTAL	**24**	**100**
Degree in MA	ELT	8	33,33
	LIT	5	20,83
	ED.SC	3	12,50
	Others	2	8,33
	None	6	25
	TOTAL	**24**	**100**
PhD Candidates in	ELT	1	4,17
	ED. SC	2	8,33
	None	21	87,50
	TOTAL	**24**	**100**
Experience in	1-4 years	1	4,17
• **teaching**	5-10 years	10	41,67
	Over 10 years	13	54,17
	TOTAL	**24**	**100**
• **the institution**	1-4 years	12	50
	Over 5 years	12	50
	TOTAL	**24**	**100**
• **marking**	1-4 years	3	12,50
	5-10 years	13	54,17
	Over 10 years	8	33,33
	TOTAL	**24**	**100**
Rater Training	YES	8	33,33
	NO	16	66,67
	TOTAL	**24**	**100**

Three of the non-native speakers were still working toward their PhDs, which were all related to English language. Participants' teaching experience ranged from 4 to 20 years and all but one had more than five years of teaching experience. Most of them worked at the current institution more than six years (*min*. 2 years and *max*. 7 years). Only eight of them mentioned that they had taken a training program or a workshop related to rater training. Table 3.1 gives a detailed depiction of the participant instructors.

3.3. Data Collection Instruments, Design and Procedures

The study was carried out to identify whether standardisation sessions conducted before writing assessments had an effect on intra- and inter-rater reliability. Therefore, the assessments of the instructors of the writing section in the proficiency exam in September were compared to the reassessments of the same papers by the same instructors in May.

The study consisted of two phases. On 12th September, 2018, the EPE writing exam was conducted. After the exam, the instructors were given a standardisation session, and directly after the session, they marked the papers. Namely, the instructors went through an actual standardisation session and marked actual EPE writing exam papers. Each instructor was allocated 22 papers in total. The names of the students were visible to the instructors. Later, each paper was double marked by the instructors who were not allowed to see the first markers' grades. Once double marking was completed, grades were compared between the instructors and those

papers with a five or more discrepancy gap were remarked and given final scores by the Discrepancy Committee. Based on the agreed grades of the instructors for the writing exam, together with grades obtained from the other sections in the proficiency exam, students either were placed into the appropriate language levels in ELS or started in their departments.

In the second phase of the study, after an eight-month interval in May, 10 student papers out of the same 22 papers graded in the post-test in September were randomly picked and given to the same instructors who had assessed them before.

Table 3.2

Data Collection Procedure

Data Collection Procedure	
September, 2018	✓ The EPE was conducted to the students. ✓ Instructors were given a writing standardisation session. ✓ Each instructor marked 22 papers. ✓ Students were allocated in an appropriate language level in ELS or started in their faculty based on the cumulative grades they obtained in every section in the exam.
May, 2019	✓ The same rubric and 10 randomly-picked EPE writing papers were delivered to the participant instructors. ✓ Instructors marked the papers and data collection procedure ended.
June, 2019	✓ Data analysis process started.

This time, to ensure the fairness of grading, the names of the students were concealed by the researcher in case some students happened to be the raters' students during the semesters. Instructors were provided with the same rubric that they had used in September. A *WhatsApp* group was formed in order to answer teachers' questions related to the marking. The reason for staying in contact with the teachers was to guide the instructors while grading, unlike the study by Diederich et. al in 1961, which was criticised by many in the field for assessors' being left on their own while they were marking. Table 3.2 presents the data collection procedure to make the process clearer.

3.4. Data Analysis

The data obtained for the study were analysed quantitatively and they were entered into the Statistical Package for Social Sciences (SPSS, version 23). To determine whether a parametric or non-parametric analysis was needed, first, a normality test was run. The normality test result showed that the data did not follow a normal distribution (see Table 3.3). Therefore, the non-parametric Wilcoxon Signed Ranks Test was run to show if there was a significant difference between the two dependent variables, which are the participant teachers' assigned scores in the pre- and post-assessments.

Table 3.3

Normality Test Results

	Kolmogorov-Smirnov[a]		Shapiro-Wilk	
Pairs	*df.*	*p.*	*df.*	*p.*
1. Assessment	240	,000	240	,000
2. Assessment	240	,003	240	,000

$p < .05$

3.5. Research Ethics

The current study was approved by the Human Subjects Ethic Committee of Middle East Technical University (METU) and it followed all required ethical considerations (Appendix C). All 24 teachers participated voluntarily in the study and signed the consent form (Appendix A). The identities of the participants were kept confidential and not shared with third parties. As stated in the consent form, there were no known risks associated with the study and teachers were free to withdraw at any time if they did not want to continue. Finally, the results were shared with participants by the researcher.

CHAPTER 4

RESULTS AND DISCUSSION

4.0 Introduction

The aim of the current study was to investigate whether standardisation sessions conducted before writing assessments had an effect on rater reliability. The study seeks answer the two following research questions:

1. To what extent do standardisation sessions conducted before writing assessments affect *intra-rater reliability*?

2. To what extent do standardisation sessions conducted before writing assessments affect *inter-rater reliability*?

The study was conducted with 24 English Language Instructors (native speakers n=4; non-native speakers n= 20) who volunteered to take part. In order to answer the research questions, each participant instructor was asked to assess proficiency exam writing section papers, once with a standardisation session in September, and once without it in May. There was an eight-month interval between pre- and post-test to avoid any recall effect. In September, each instructor assessed 22 papers, and in May, 10 student papers were picked randomly for each instructor among them for re-assessment. As a result, 24 instructors assessed the same 10

papers twice so the number of the re-evaluated papers in the study was 240 in total (24x10=240).

This chapter consists of two main sections. In the first section, data gathered from participant instructors' pre-assessments in September, and post-assessments in May, will be analysed quantitatively and compared. In the second section, research questions will be discussed in line with findings and related literature. They will be given in accordance with the research questions, respectively.

4.1 Results and Findings

4.1.1 The Effect of Standardisation Sessions on Intra-Rater Reliability

Table 4.1 illustrates the descriptive statistics of the scores for two different assessments of the participant instructors. Accordingly, the mean scores of the first assessment (*M*=17, 77) are higher than in the second assessment (*M*=16, 89). The minimum score given in the first assessment is 5 while the top score is 25. Similarly, the highest score given in the second assessment is 25 and the lowest score is zero.

Table 4.1

Descriptive Statistics of Scorings

Assessments	*min*	*max*	*M*	*SD*
1	5	25	17,77	0,307
2	0	25	16,89	0,299

Based on the normality test results, the non-parametric Wilcoxon Signed Ranks Test was conducted to determine to what extent standardisation sessions held before writing assessments affect intra-rater reliability (see Table 4.2). As aforementioned, each of the 24 instructors was awarded 10 papers for a second assessment after an eight-month interval. Therefore, as seen in Table 4.2, the mean scores of the second assessments of 17 instructors – Instructors #1, #2, #3, #4, #7, #8, #9, #10, #11, #12, #14, #15, #16, #17, #18, #19 and #22 – were lower than their first assessments' scores whereas the mean scores of the second assessments of 7 instructors – Instructors #5, #6, #13, #20, #21, #23 and #24 – were higher than their first assessments.

Analysis of the Wilcoxon Signed Ranks Test indicated that the difference between the first and the second assessments of Instructor #1 (z = -2,349, p= .019,), Instructor #2 (z =,-2, 31, p = .0210), Instructor #7 (z = -2, 655, p = .008), Instructor #9 (z = -2,315, p = .021), Instructor #11 (z = - 2, 453, p = .014) Instructor #12 (z = -2, 668, p = .008), Instructor #13 (z = -2, 209, p = .027), Instructor #15 (z = -2,836, p = .005), Instructor #19 (z = -2,222, p = .026) and Instructor #24 (z = -2, 245, p = .025) were statically significant whereas no significant difference was seen in the other 14 participants.

According to Table 4.2, while the second assessments of Instructors #1, #2, #7, #9, #11, #12, #15 and #19 were significantly lower, Instructors #13 and #24 were significantly higher in their second scorings.

Table 4.2

Intra-Rater Reliability Test Results

Instructors	Papers *n*	1. Assessment in September		2. Assessment in May		Difference	Z	p
		M	***SD***	***M***	***SD***			
# 1	10	18,90	2,85	15,60	4,03	-3,30	-2,349	**0,019***
#2	10	20,10	2,42	17,50	2,37	-2,60	-2,31	**0,021***
#3	10	16,60	3,41	15,50	3,92	-1,10	-0,562	0,574
# 4	10	23,00	2,31	21,40	3,17	-1,60	-1,632	0,103
#5	10	17,70	2,45	18,30	3,80	0,60	-0,052	0,959

Table 4.2 (continued)

#6	10	16,30	3,74	17,30	3,62	1,00	-1,103	0,270
#7	10	19,30	3,59	16,00	2,54	-3,30	-2,655	**0,008***
#8	10	20,10	3,21	20,00	3,46	-0,10	-0,577	0,564
#9	10	20,90	4,43	17,30	3,77	-3,60	-2,315	**0,021***
#10	10	20,10	3,81	18,90	3,00	-1,20	-1,845	0,065
#11	10	18,30	2,41	15,30	3,33	-3,00	-2,453	**0,014***
#12	10	19,80	3,49	15,00	2,79	-4,80	-2,668	**0,008***
#13	10	16,60	6,62	18,40	5,50	1,80	-2,209	**0,027***

Table 4.2 (continued)

#14	10	17,70	5,50	16,50	7,34	-1,20	-0,841	0,400
#15	10	17,50	3,87	14,50	4,03	-3,00	-2,836	**0,005***
#16	10	17,70	6,24	16,90	5,74	-0,80	-1,318	0,187
#17	10	21,40	3,10	20,40	3,57	-1,00	-0,846	0,397
#18	10	21,20	3,39	21,10	2,42	-0,10	-0,051	0,959
#19	10	15,70	4,27	12,80	4,10	-2,90	-2,222	**0,026***
#20	10	16,10	3,41	17,90	5, 65	1,80	-1,489	0,137

Table 4.2 (continued)

#21	10	13,10	4,72	15,40	4,60	2,30	-1,944	0,052
#22	10	13,40	5,10	13,00	4,88	-0,40	-0,352	0,725
#23	10	12,00	3,94	13,10	5,22	1,10	-0,655	0,512
#24	10	13,00	3,97	17,20	3,36	4,20	-2,245	**0,025***

*: $p < .05$

The most significant difference between the first and second assessment belonged to Instructor 12 and Instructor 24 with a difference of -4.80 and 4.20 respectively.

Table 4.3 shows the total number of papers graded twice by the instructors once with and once without a standardisation session, and the mean scores of their first and second assessments. Accordingly, while the mean scores of the total 240 writing papers graded by 24 instructors in the first assessment in September was 17.77, they decreased to 16.89 in the second assessment in May.

Table 4.3

Comparison between the Pre- and Post-test for All the Papers

Assessment	*Papers n*	*M*	*SD*	Difference	Z	p
#1	120	17,77	4,761	-0,88	-3,604	**0,000***
#2	120	16,89	4,631			
TOTAL	240					

*: p< .05

Based on the results of the Wilcoxon Signed Ranks Test, the difference between the first and second assessments was statistically meaningful (z= -3,604, p= .000). As seen clearly in Table 4.3, the second assessments of the instructors were significantly lower than their first ones.

4.1.2 The Effect of Standardisation Sessions on Inter-Rater Reliability

As mentioned before, every student paper (n=120) was graded by two instructors. With 24 participant teachers in the study, each paper was graded by 12 pairs. With regard to the second research question, the aim of this section is to identify whether or not standardisation sessions have an effect on inter-rater reliability. In other words, this section of the study seeks to discern to what extent standardisation sessions contribute to partner markers' consistency. The Wilcoxon Signed Ranks Test was conducted to see if the difference between the first and the second assessment of each pair was significant.

Table 4.4 shows the differences between the pairs' (1) first scores, (2) second scores, and (3) first and second scores. When the differences between the first and second assessments of each pair were compared, it was found out that the mean scores of 8 Pairs (Pairs # 1, #3, #4, #5, #8, #9, #10 and #12) decreased while the mean scores of Pairs #2, #6, #7 and #11 increased in the second assessment.

According to the Wilcoxon Signed Rank Test, while the difference between the first and the second assessments of Pair #1 (z = -2,349 p = .019), Pair #4 (z = -2, 655, p = .008), Pair #5 (z = -2, 315, p = .021), Pair #6 (z = -2,453, p = .014), Pair #7 (z = - 2,209, p = .027) Pair #8 (z = - 2,836, p = .005) and Pair #10 (z = - 2,222, p = .026) were significant, the difference between other five pairs were not. The second

Table 4.4

Inter-Rater Reliability Test Results

Pairs	Papers	Difference in 1.Assessment		Difference in 2. Assessment		Difference	Z	p
	n	***M***	***SD***	***M***	***SD***			
# 1	20	-1,20	2,20	-1,90	4,04	-0,70	-2,349	**0,019***
# 2	20	-6,40	3,06	-5,90	2,88	0,50	-0,562	0,574
# 3	20	1,40	4,62	1,00	4,69	-0,40	-0,052	0,959
# 4	20	-0,80	2,30	-4,00	1,76	-3,20	-2,655	**0,008***
# 5	20	0,80	2,70	-1,60	2,07	-2,40	-2,315	**0,021***
# 6	20	-1,50	2,01	0,30	3,56	1,80	-2,453	**0,014***

Table 4.4 (continued)

# 7	20	-1,10	2,69	1,90	4,58	3,00	-2,209	**0,027***
# 8	20	-0,20	3,39	-2,40	2,84	-2,20	-2,836	**0,005***
# 9	20	0,20	1,03	-0,70	3,06	-0,90	-0,846	0,397
#10	20	-0,40	1,51	-5,10	4,84	-4,70	-2,222	**0,026***
#11	20	-0,30	2,75	2,40	2,72	2,70	-1,944	0,052
#12	20	-1,00	1,76	-4,10	6,12	-3,10	-0,655	0,512

*: $p < .05$

assessments of Pairs #1, #4, #5, #8, and #10 were significantly lower whereas Pairs #6 and #7 scored higher in the pre-test. The most significant difference belonged to Pair #10 and Pair #4, with a difference of -4.70 and -3.20, respectively.

4.2 Discussion

4.2.1 The Effect of Standardisation Sessions on Intra-Rater Reliability

The result of the data analysis revealed that out of 24 participant instructors who graded the same papers twice after an eight-month interval, there was a significant difference between 10 instructors' pre- and post-test mean scores. Whereas the second assessments of Instructors #1, #2, #7, #9, #11, #12, #15 and #19 were significantly lower, Instructors #13 and #24 were significantly higher in their second scorings (see Table 4.2). However, although there was a statistical difference for only 10 instructors' (almost 42%) two different markings, when instructors' mean scores were individually compared, it became obvious that the mean scores of each instructor in the post-test were different from the first one (see Table 4.2). While seventeen instructors assigned lower grades, seven of them scored higher in their second assessments. When the scoring in the rubric was taken into consideration, the minimum difference between the two scorings for a paper could be as low as 1 point. That means if a paper was assigned a different score in the post-test, it was scored at least 1 point lower or higher. For instance, if an instructor scored a paper 19 in the pre-test, it meant he gave a minimum of 18 or 20 in the post-test.

On the one hand, as writing assessment is a qualitative testing method and based on raters' subjective judgements, it is normal to see differences between raters'

first and second scorings. In other words, it was not surprising to see changes in the instructors' first and second grades because it was understandable that their judgement could have changed to some extent in time. On the other hand, when the importance of proficiency exams is taken into consideration, even one point could lead a student to pass or fail. Especially, if a student is a repeat student, even one point may cause him to start his department or repeat the whole level he has already studied at and failed (Tanrıverdi-Köksal, 2013, p.109). Therefore, even one point could make a difference for the test taker although it might not seem such a large difference (Myford & Wolfe, 2000, cited in Tanrıverdi-Köksal, 2013). When the first and second scores of the papers were studied, it was found that, out of 240, only 30 papers' first and second scores matched while the rest were either higher or lower. It means the average proficiency exam scores of most students might have changed if the scores of the second assessment had been counted as the students' proficiency writing exam scores.

To be more precise, when each instructor is studied based on the data in Table 4.2, it is obvious that the most consistent rater is Instructor #8 with -1,10 difference between his two marking whereas the most inconsistent one is Instructor #12 with a difference of -4.80. Consequently, proficiency exam scores of many students graded by Instructor #12 would most probably have been affected to a great extent if his second grading had been recorded as the students' proficiency exam scores.

However, as what directly affects students' scores are not individual scores but a combination of two partners' marks, this part was further discussed in the

following section for a better analysis after scores of both raters were checked and verified.

Furthermore, when Table 4.2 is studied, it is obvious that most instructors (*n*= 17) graded more severely in their second assessments. In other words, when instructors did not receive a standardisation session before grading, they had harsher judgements. Therefore, it could be suggested that standardisation sessions might have a positive effect on reducing severity to a great extent, though not eliminating completely. This finding was in line with the view of Greatorex, 2002. Moreover, although instructors used the same rubric in both sessions, their two scorings showed difference, usually more severely in the second one. Therefore, although there was significant literature about the importance of rubric use, supporting its efficiency in (1) increasing valid judgement of the assessor and, therefore, reliability (Jonsson & Svingby, 2007; Kohn, 2006 and Silvestri & Oescher, 2006); and (2) rendering better consistency for intra- and inter-rater reliability (Moskal & Leyden, 2000), the finding of the current study fit with the views of many in the field (Chapman & Inman, 2009; Dawson 2009; Gonzales, Trejo & Roux 2017; Kohn, 2006; Reddy & Andrade, 2010 and Rezai & Lovorn, 2010) who underlined that a rubric itself could not provide better reliability but should be backed-up with rater training, or rather standardisation sessions.

The conclusion that standardisation sessions make a difference in raters' judgements was also supported when two different scores of the instructors were individually analysed (see Table 4.2). It was observed that only 12.5 % of the scores did not change whereas 87.5 % of them were graded either lower or higher. Besides, as clearly seen in Table 4.3, there is a significant difference when the mean scores of

the pre- and the post-test of all the papers are compared. The mean scores of the second assessments (M=17, 69) are significantly lower than the first assessments (M=16, 89), which supports the finding that standardisation sessions affect raters' assessments.

There might be some reasons for the differences between the two markings of the instructors. First, as mentioned above, standardisation sessions could be beneficial to improve and validate the assessment performance of raters to a great extent, affecting intra-rater reliability positively. As McNamara has suggested, the main purpose of these sessions are to help raters be internally consistent in their own evaluations. This will, in the end, result in higher inter-rater reliability (1996). However, together with or apart from the effect of standardisation sessions, there might be some other reasons for the difference. For instance, the reason why most instructors graded the papers more severely in the post-test could be that they knew their harsh scorings would not cause any students to fail in the end. Because the pre-test was an actual examination, teachers might have scored papers more leniently to favour the students. Moreover, in the first assessment, the names of the students were visible to the instructors. Owing to the fact that repeat students also took the proficiency exam in September, there might have been some students familiar to the instructors and this might have caused more tolerant markings for some students. However, in the post-test, the researcher hid the names of the students on the paper, just in case instructors happened to recognize some students and mark them accordingly. Hence, this could have led to harsher discretion from the instructors. Last but not least, another reason for the difference might be that, in the pre-test, instructors had to grade 22 papers at one sitting after a long standardisation session

so they might have felt exhausted. They may have rushed to finish marking as soon as possible. However, in the post-test, because of the hectic and different schedule of each participant, the researcher could not gather instructors in a classroom and force them to finish marking in the same day. Therefore, not having the same grading conditions might have led to differences between the two marking sessions. Moreover, instructors were awarded only 10 papers for the reassessment, not 22 as in the pre-test. This may have also led instructors to analyse the papers more profoundly and to mark more severely. Simply put, most instructors (*n*= 17) graded more severely in their second assessments, in which they had to grade fewer papers. This finding was in contrast to Pinot de Moira et. al (2002) who suggested that the more paper raters scored at one sitting, the severer they became.

Keeping in mind that each and every instructor's mean scores showed differences to some extent, it should be noted that only 10 raters had a significant difference in the data analysis. To be more precise, while almost 42% of the instructors (*n*=10) showed meaningful significance, more than 50% percent (*n*=14) did not. There might be some reasons for this. First of all, although this finding seemed to support the literature (Baird et. al., 2004; Black, 1962; Greatorex et. al., 2002; Pufpaff, Clarke & Jones, 2015 and Shaw, 2002) that suggested standardisation sessions might not be as effective in rater reliability, the researcher wants to underscore the importance of these sessions' efficiency as well their effectiveness (Pufpaff, Clarke & Jones, 2015). Namely, throughout their hectic work schedule, instructors might not take advantage of the session adequately enough. As a result, the standardisation session might not be as efficient and effective on these instructors' judgements as anticipated. However, even though there was no

statistically significant difference for these 14 instructors, the researcher needs to emphasize once more that their two scorings still showed variety, which, as aforementioned, could make a significant difference for the test taker. To show how much this could affect a test taker's final score, one of the papers (paper #7) marked by Instructor #14, who did not show significant difference in the data analysis, was studied. When all the papers were compared individually, this paper showed the largest difference between its first and second scoring. In his first assessment, the instructor scored the paper 12, while in the second one he scored the same paper zero, checking the band that reads 'The text is unrelated to the topic. The student fails to answer the question implicitly or explicitly.'

When the researcher checked paper #7, she agreed with the instructor that the paper was off-topic and it did not answer the question given in the prompt. Seeking a second opinion, the researcher showed the paper to the Head of Testing Unit. The Head also confirmed that the paper was off-topic. Obviously, although Instructor #14 did not demonstrate a statistically significant difference in the data analysis, if his second score for this paper had been recorded as the exam grade, the student's proficiency exam score would have changed drastically.

More interestingly, the second judgement of the instructor, which he made with the absence of a standardisation session, seemed more appropriate than his first one after he received a session. Based on this finding, it might be possible to say that standardisation sessions do not always lead to better judgement. However, out of 120 student papers, there was only one paper from which such a conclusion could be drawn. Moreover, when the other partner's (Instructor #13) scoring for the same paper was checked, he scored the same paper 9 in the pre-test and 12 in the post-test.

Obviously, without the absence of a standardisation session, he scored higher although the paper was confirmed to be off-topic. Therefore, generalizing that standardisation sessions might lead to faulty judgement could be inaccurate. However, identifying the reasons behind this huge discrepancy via an interview with Instructor #13 and Instructor #14 could have led to a better conclusion.

- *Analysis in Relation to Rater Differences*

A sizeable amount of literature has focused on whether or not rater difference is a factor that affects raters' judgement. Experience in grading and language background are some of these differences mostly discussed in this literature (Barkaoui 2007; Black, 1998; Brown 2003; Davies 2008; Gonzalez & Roux, 2013; Hamp-Lyons & Davies 2008; He, Gou, Chien, Chen & Chang 2013; Lim, 2011; Lumley 2006; McNamara 1996; Shi 2001 and Shi, Wan & Wen, 2003).

In the study, the Instructors #3, #6, #7 and #9 were native speakers of English. Based on the findings (see Table 4.2), only Instructors #7 and #9 showed significant discrepancies while the other two did not. This finding itself might not make sense because of the limited number of native speaker participants in the study (n=4). Therefore, these data should be studied together with non-native speaker participants who also showed a significant difference (Instructors #1, #2, #11, #12, #13, #15, #19 and #24) in the data analysis. Accordingly, out of four native instructors, two of them showed a significant difference while out of twenty non-native speaker instructors, eight of them scored significantly different in the post-test. Namely, 50% of native and 40% of non-native participants demonstrated a significant difference. Given that both parties shared such a close percentage, it could be concluded that ratings of native and non-native speakers were not that different

from each other. This was contrary to the findings of Barnwell, 1989; Fayer and Krasinski, 1987; and Galloway, 1980 who concluded that native and non-native speakers' judgements showed difference; but consistent with Brown, 1995; Hill, 1996; Kim, 2009 and Zhang & Elder, 2011. Furthermore, the findings showed that out of ten instructors showing statistically meaningful differences, Instructors #1, #2, #7, #9, #11, #12, # 15 and #19 scored significantly lower while Instructors #13 and #24 scored significantly higher in the post-test. Knowing that Instructors #7 and #9 were native speakers and Instructors #13 and #24 were non-native speakers of English, this finding is at variance with Kang (2008), who concluded non-native speakers had a tendency to mark more severely. However, a future study with more native speaker participants would probably yield more accurate results. Moreover, as in Shi (2001), native and non-native speakers could be asked to justify their scorings through think-aloud protocol during their grading to be able to better analyse and compare two groups' understanding of writing assessment.

Furthermore, Instructors #3, #6 and #7 were lay raters, whose experiences in rating were 2.5, 2 and 3 years respectively and only Instructor #7 scored significantly lower in the post test. However, considering that the other nine experienced instructors (Instructors #1, #2, #7, #11, #12, #15, and #19 marking significantly lower and Instructors #13 and #24 significantly higher) and other novice raters (Instructors #3 and #6) who did not show a meaningful difference, it could be suggested that regardless of experience in grading, standardisation sessions had an effect on raters' scorings (Shohamy, Gordon & Kraemer, 1992 and Weigle 1999). Therefore, the study is compatible with the suggestion of Shohamy, Gordon and

Kraemer that decision makers should not be worried about raters' background but rather a proper rater training (1992, p. 31).

4.2.2 *The Effect of Standardisation Sessions on Inter-Rater Reliability*

As multiple judgments lead to a truer final score than any single judgment (Hamp-Lyons, 1990), for a more reliable grading, double marking procedure is followed in the university where the study took place. Once teachers finish scoring one group of students, they reassess another group of student papers as a second marker, without seeing the first markers' grades. The exam score totalled up to 25 and the allowed discrepancy between two markers was a maximum of 4. The cases of extreme discrepancies, such as 5 and more, required another evaluation by the Discrepancy Committee. If the markers' grades were within the allowed discrepancy (1-4), then a combination of two separate marks was the student's final score.

In the current study, each student paper (n=120) was also double-marked both in the first and second assessment. Therefore, the same paper was marked twice so, in total, 240 papers were re-evaluated (120x2=240). Participant instructors formed twelve pairs. Pairs were not formed by the researcher but by the Head of Testing Unit randomly because the pre-test that took place in September was the actual proficiency exam. Hence, to be able to see the inter-rater consistency, the researcher kept the pairs in the May post-test as they were in the pre-test. Each consecutive instructor formed one pair, i.e. Instructor #1 and Instructor #2 were Pair #1 while Instructor #3 and Instructor #4 were Pair #2.

The data analysis demonstrated that out of 12 pairs, 7 of them showed significant difference (Pairs #1, #4, #5, #6, #7, #8 and #10). Whereas Pairs #1, #4,

#5, #8 and #10 were significantly lower, Pair #6 and #7 scored meaningfully higher in the post-test (see Table 4.4).

As seen in the table, almost 60% of the raters scored less consistently in the post-test, in which they did not receive a standardisation session. The most significant difference belonged to Pair #10 with a difference of -4.70. To be more precise, based on the data in the table, the discrepancy between the instructors' scores was only 4 in the pre-test but they increased to 47 in the second assessment, which was a giant leap. Therefore, Pair # 10 was the least consistent among all the pairs. They were followed by Pair #4 with a difference of -3.20.

Moreover, although slightly more than 42% of the pairs did not show a significant difference in the data analysis, comparison of the pairs' first and second assessments showed that each pair scored differently in the post-test (see Table 4.4). As literature on assessment maintains, it is normal to spot differences between the first and second assessor because assessment of direct methods is based on subjective judgments of raters. However, when each pair was individually studied, it was found out that the differences between the pairs, even those who did not a show significant difference in the data analysis, would drastically change the students' exam scores. For instance, although Pairs #11 and #12 were not statistically different in the data analysis, they scored quite differently in two assessments. Namely, while the difference between the scores of the instructors forming Pair #11 was 3 in the pre-test, it increased to 24 in the post-test. Likewise, in the first assessment, the difference between the scores of the raters forming Pair #12 was 10 but it increased to 41 in the second one. Variations to some extent are acceptable but raters who consistently rate too high or too low should adjust their standards (Knoch, 2009, p.

31). Therefore, spotting such large discrepancies for many papers graded without a standardisation session, the study concluded it is of utmost importance that raters receive a standardisation session before grading. This echoes the suggestion of Greatorex and Shannon that 'a standardisation session ensures; (1) each assessor to consistently make valid decisions, (2) all assessors to make the same decision on the same evidence base, and (3) all candidates to be assessed fairly' (2003, p. 3).

Together with this finding, the papers whose discrepancies were more than 4 were also spotted. Table 4.5 illustrates the number of papers that each pair scored beyond the allowed discrepancy gap in the post-test.

Table 4.5

Papers beyond Allowed Discrepancy Gap in the Second Assessment

Pairs	Paper Sets Graded by Each Pair	Papers Graded beyond Allowed Discrepancy Gap
	n	n
# 1	10	3
# 2	10	3
# 3	10	4
# 4	10	6
# 5	10	0
# 6	10	2
# 7	10	2
# 8	10	2
# 9	10	1
# 10	10	4
# 11	10	3
# 12	10	5
TOTAL	120	TOTAL 35

In the current study, each pair graded 10 paper sets twice. Therefore, in total, 120 paper sets were reassessed (12x10=120). As seen in the table, out of 120 paper sets, 35 papers in total are not marked within the allowed discrepancy gap. Out of ten paper sets that Pair #4 graded, they scored six papers beyond the allowed discrepancy gap in their second assessments. It is followed by Pair #12 with five papers. However, when the first grades of the pairs were checked, there were no discrepancies between the pairs for these papers. Therefore, the study found out that standardisation sessions were 'effective in terms of bringing participant teachers' judgements in line with the rest in terms of both marks and the procedures by which they arrived at those marks' (Weigle, 1994 cited in Meadows & Billington, p. 51). Because the aim of these sessions is to, (1) reconcile teachers' judgements as much as possible by providing them with clear understanding of the rubric and the criteria that teachers need to take into consideration during their assessment; (2) to familiarize raters with the scoring procedures; and (3) prepare them for dealing with the unexpected situations during rating (McNamara, 1996, p. 92), it was not surprising to see better consistency between pairs' pre-test scores given right after the standardisation session.

Moreover, though standardisation sessions are not considered time-efficient, the study suggests the just opposite. As seen in Table 4.5, 60% of the paper sets rescored by Pair #4, and 50% of the paper sets reassessed by Pair #12 were not within the allowed discrepancy gap. Therefore, these papers would have needed to be assessed one more time by the Discrepancy Committee if this had been the actual marking session for the proficiency exam. More importantly, out of 120 papers, 35 of them were graded beyond the allowed discrepancy gap, which means they were

supposed to be assessed by the Committee members. In September 2018, 413 students took this proficiency writing exam, and out of 413 student papers, only 120 were included in the present study. Now that out of 120 papers, 35 papers were to be rechecked by the Committee members, out of 413 papers, around 120 papers would have required another reassessment by the Committee if it had been the actual grading. Seeing that the university where the study was carried out had a relatively small student population, the amount of papers requiring a third evaluation would increase noticeably in more crowded universities. Therefore, it is obvious that standardisation sessions are time-efficient and they ease grading.

Another interesting finding was seen in the discrepancies between the scores of pairs with the absence of a standardisation session although they used the same rubric in both marking sessions. Therefore, the study showed that using a rubric *per se* did not help increase consistency between pairs. As a result, 'deliberate training and agreement of raters before any [assessment] process seemed needed' (Kayapınar, 2014, p. 127). This finding was also compatible with previous studies suggesting that using a rubric did not work to improve rater reliability *per se* (Gonzales, Trejo & Roux 2017; Kondo, 2010 and Lovorn & Rezai 2011).

- *Instructor #12*

In the data analysis of intra-rater reliability, Instructor #12 showed the most significant difference between his first and second marking. He scored significantly lower in the post-test than in the pre-test. However, as what directly affects students' scores are not individual scores but a combination of two partners' marks, in this section, his scores were studied together with his partner's reassessment. In this way, it was discovered whether or not –and, if any, to what extent- his inconsistency

caused any difference in students' final scores. Instructor #12 was paired with Instructor #11 and they formed Pair #6.

Table 4.6

Comparison of the Two Assessments of Instructor 11 and Instructor 12

	1. Assessment Scores *M*	2.Assessment Scores *M*	Difference
Instructor 11	18, 30	15, 30	- 3, 33
Instructor 12	19, 80	15, 00	- 4, 80

As clearly seen in Table 4.6, both partners scored much more severely in their second assessments. As illustrated in Table 4.2, Instructor #11 was also one of the raters that showed a significant difference between her two assessments, as did her partner (z= -2,453, p= .014). While Instructor #12 was the rater who showed the most significant difference in his two scorings, Instructor #11 was the fifth in rank out of 24 participants. Therefore, she was also one of those whose two assessments were inconsistent with the absence of a standardisation session.

For a better picture, Table 4.7 shows the students' final scores based on the first and the second assessments of Pair #6. When the final scores of the two assessments were compared, it was detected that half of them showed large discrepancies. The first and the second assessments of papers #4 and #9 especially showed the biggest difference between their two assessments (8 for both paper #4 and paper #9).

Moreover, the discrepancies between the pairs for paper #1 and paper #8 in the second assessment exceeded the allowed discrepancy gap so these papers had to

Table 4.7

Comparison of Final Scores of the Papers Based on the Two Assessments of Instructor 11 and Instructor 12

Papers	1. Assessment Scores			2. Assessment Scores		
	Instructor 11	Instructor 12	Final Score	Instructor 11	Instructor 12	Final Score
#1	18	16	17	17	12	The Committee
#2	21	25	23	17	19	18
#3	18	19	19	14	17	16
#4	19	23	21	14	12	13
#5	15	16	16	14	16	15
#6	20	19	20	22	18	20
#7	14	15	15	11	11	11
#8	17	20	19	11	17	The Committee
#9	21	24	23	15	14	15
#10	20	21	21	18	14	16

be re-evaluated by the Committee. Briefly, (1) if the second assessment had been the actual exam, two papers out of ten would require a third eye to reassess; and (2) the second scorings might have caused almost half of the students to fail. To be more precise, Students #2, #4 and #9 would receive much lower grades. If these were borderline students, they might have failed if the second scoring had been their actual exam scores. Considering that proficiency exams are high stake exams, even one point could make a great difference for test takers. Seeing that there were no discrepancies for ten papers graded in the first assessment after the standardisation session, the study concluded that standardisation sessions constituted an important part of direct writing assessment.

- *Paper #7 Graded by Instructor #13 and Instructor #14*

In his second assessment, Instructor #14 graded a paper (paper #7) zero, checking a band that reads 'The text is unrelated to the topic. The student fails to answer the question implicitly or explicitly.' However, he had scored the paper 12 in the pre-test. Therefore, this paper showed the largest discrepancy among all the papers although the instructor's mean scores from his two different assessments did not show a significant difference in the data analysis.

Because students' final scores are a combination of two assessors' marks, in this section, the differences between the instructors who graded this paper were studied. Instructor #13 and Instructor #14 graded this paper and they formed Pair #7. Table 4.8 shows the differences between the pair's first and second assessments for this paper.

Based on the data given in the table, the final score of the paper is 11 in the pre-test. However, for the second assessment of the pairs, the Committee's verdict is needed. As explained in detail in the first section of this chapter, the decision given by Instructor #14 that the paper was off-topic was confirmed by the Testing Head. It means the second assessment of Instructor #14, given with the absence of a standardisation session, was correct while it was faulty for Instructor #13. However, both instructors scored an off-topic paper 9 and 12 in their first assessments. The reason for this might be that teachers could be marking the papers more leniently in the actual exams. Moreover, having to mark twenty-two papers at one sitting after a long standardisation session in the pre-test might have negatively affected instructors' morale and judgements. Therefore, the efficiency of standardisation sessions, as well as their effectiveness on rater consistency, should also be studied in the future. Moreover, why Instructor #13 gave 12 in his second assessment –which was higher than his first scoring- and why Instructor #14 scored 12 in his first assessment and changed his mind in the second one could also be investigated via a think-aloud protocol and/or interviews for a better understanding of these kinds of judgements.

Table 4.8

Scores of Paper #7 Graded by Instructor 13 and Instructor 14

	Instructor 13		Instructor 14	
Paper	1. Assessment	2. Assessment	1. Assessment	2. Assessment
#7	9	12	12	0

- *Pairs Who Showed No Significant Difference in the Second Assessment*

As seen in Table 4.4, there are five pairs who did not show a significant difference in their second assessments (Pairs #2, #3, #9, #11 and #12). However, as seen in Table 4.5, all these pairs still have papers that should be reassessed by the Committee (Pairs #2 and #11 three papers, Pair #3 four papers, Pair #9 one paper, and Pair #12 five papers). Moreover, when the instructors who formed these pairs were studied on a case by case basis, it was learned that Instructor #24, one of the markers of Pair #12, already showed significant difference in the data analysis for intra-rater reliability. He was actually the second in rank out of 24 participants.

Last but not least, although Pairs #11 and #12 showed no significant difference in the data analysis, the discrepancies between their two scorings were still high. The difference between the scores of Instructors #21 and #22, who formed Pair #11, was 2.7 while it was 3.1 for Instructors #23 and #24 that formed Pair #12 (see Table 4.4).

Briefly, when all the pairs were examined one by one, it was observed that most students' final scores would have been affected greatly if the second grading had been the students' actual exam scores. Therefore, it can be concluded that standardisation sessions are effective in raters' evaluations.

CHAPTER 5

CONCLUSION

5.0 Introduction

The purpose of the current study was to investigate whether or not –and, if any, to what extent- standardisation sessions were effective on rater reliability. To this end, the study addressed two research questions:

1. To what extent do standardisation sessions conducted before writing assessments affect *intra-rater reliability*?

2. To what extent do standardisation sessions conducted before writing assessments affect *inter-rater reliability*?

24 English language instructors that worked in the preparatory school of a foundation university participated in the study. The study consisted of two phases: The first phase was in September, where the participant instructors graded the actual proficiency exam writing papers. Each instructor graded 22 papers. Before they started marking the papers, they were given a standardisation session in which first the rubric and the exam specifications were explained and later instructors individually marked five student papers chosen randomly by the testing office. After grading these papers, the instructors and the Testing Unit member discussed the grades given for each band so that teachers could reconcile their individual judgements as much as possible. The analytic rubric used for the grading had been

developed by the Testing Unit members and then approved by the Testing Head and the ELS administration respectively. It consisted of four different bands. (1) Content and Organization out of 10; (2), Coherence out of 5; (3) Lexical out of 5, and (4) Grammar out of 5 points. The minimum score that the students could obtain for each band was 1. Therefore, the lowest grade a student could get from the exam was 5 while the maximum score was 25. However, if the paper was off-topic, it would get a zero.

Right after the standardisation session, the instructors started to grade the student papers. The names of the students were visible to the instructors. They were supposed to finish marking by the end of the day. Each paper was double-marked. First, one instructor finished one pack of papers assigned to them. Later, without seeing the grades of the first marker, a second marker graded the same papers. The final grades of the papers were a combination of the two separate marks. The papers with 5 and more discrepancy gap were remarked by the Discrepancy Committee whose members were Testing Unit teachers.

In the second phase of the study which took part after an eight-month interval in May, 10 papers were randomly chosen by the researcher for each instructor. This time, to ensure the fairness of grading, the names of the students were hidden by the researcher in case some students happened to be the raters' students during the semesters. Like in the first phase, each paper was double-marked, so out of 24 participant instructors, there were 12 pairs in the study. Briefly, 120 students' writing exams were included in the study and graded twice by the same instructors. As a result, each instructor re-graded 10 papers (24x10=240) and each pair 20 papers (12x20=240) for the current study.

This chapter consists of four main sections. In the first section, a brief summary of the findings will be presented. In the second section, limitations of the study will be given. The third section will introduce the pedagogical implications while the final section will discuss suggestions for further research.

5.1. Conclusion

Many important names in the field have studied rater reliability. To the knowledge of the researcher, there are very few studies in the literature that have dealt with both intra- and inter-rater reliability. However, as Saxton, et. al., suggest, it is useful to 'focus both on inter- and intra-rater reliability, and examine both types of behaviour as a path to improving assessment process' (2012). Therefore, the current study differs from many others for studying the effect of standardisation sessions on both intra- and inter-rater reliability.

The findings revealed that the absence of a standardisation session affected raters' judgements to a great extent. First, all the instructors scored higher or -mostly- lower in their second assessments. Although some discrepancies could be tolerated in qualitative assessment methods, it should be noted that proficiency exams are high stake assessments and even one point may make a great difference for test takers, especially for borderline students (Myford & Wolfe, 2000 cited Tanrıverdi-Köksal in 2003). Moreover, when the mean scores of all the papers (n=240) were studied, there was a significant difference between the instructors' first and second assessments. 85% of the scores in the second assessment were different from the first one. Also, it was found out that the instructors had a tendency to grade papers much more severely when they did not attend a standardisation session.

Second, although instructors used the same rubric for both assessments, most instructors graded more harshly in the post-test. Therefore, it could be concluded in the study that rubric use *per se* does not provide high rater reliability, hence, raters should be given training about how to use it productively (Chapman & Inman, 2009; Dawson 2009; Kohn, 2006 and Reddy & Andrade, 2010).

Moreover, the study concluded that scores of native and non-native instructors were not that different from each other which was at variance with some authors like Barnwell, 1989; Fayer and Krasinski, 1987; Galloway and 1980 but consistent with Brown, 1995; Hill, 1996; Kim, 2009 and Zhang & Elder, 2011. Besides, the finding of the study was contrary to Kang (2008) who suggested that non-native speakers had a tendency to mark more severely. However, because of the limited number of native speaker raters (n=4), the findings here may not be generalized.

Furthermore, when lay and experienced raters were compared, there was no significant difference between them. Therefore, it could be suggested that, regardless of experience in rating, standardisation sessions had an effect on raters' scorings. This finding was in line with Shohamy, Gordon and Kraemer, 1992 and Weigle 1999. As Shohamy, Gordon and Kraemer suggested, 'decision makers should not be worried about raters' background but proper rater training' (1992, p. 31).

In addition to the findings above, it was discovered that, with the absence of a standardisation session, there were outstanding inconsistencies between raters. Consequently, the number of papers that was graded beyond the discrepancy gap increased. Accordingly, the study proposed that standardisation sessions were effective in reconciling raters' judgements (Weigle, 1994 cited in Meadows &

Billington, p. 51). Also, because those papers need a third marker for another reassessment, standardisation sessions could be considered time-efficient.

In short, the study concludes that standardisation sessions are effective in terms of providing both inter- and intra-rater reliability which fits the views of many (Barkhuizen, Knoch, & Randow, 2007; Brown, 1995; Elder, Lunz, Wright & Linacre, 1990; Greatorex & Shannon 2003; Knoch, 2009; Lumley & McNamara, 1995; McNamara, 1996; Robinson, 2000; Shohamy, Gordon & Kraemer,1992; Sweedler-Brown, 1985; Wang, 2010 and Weigle, 1998).

5.2 Limitations of the Study

There were several limitations of the study that should be taken into consideration. First, due to the data collection procedure in the study, instructors had to mark the same papers twice. Therefore, the main limitation of this study was that the participant instructors may not have forgotten the scores they gave to the papers in the first session so they might have been affected by their first scores. However, similar test-retest studies in the related literature gave a five-week (Tanrıverdi-Köksal, 2013) to three-month (Chaplen, 1969) interval between two tests. Therefore, considering the eight-month interval between two assessments in the current study, the researcher believes that the carry-over effect must have been eliminated to a great extent, if not completely. Nevertheless, it can still be considered one of the major limitations in the present study.

Another limitation was the number of the participants. The study was conducted with 24 English language instructors in an institution that employs more than 60 English instructors. However, because participation in the study was on a

voluntary basis and the teachers at the institution had hectic schedules, the researcher could not ask all the teachers to participate in the study. Moreover, because the participant instructors had always attended a standardisation session before each writing assessment in ELS, they were already aware of what was expected from them in their second assessment, in which they did not attend a standardisation session. Therefore, their knowledge might have affected their judgement and scores.

Not providing the same marking conditions in the post-test as in the pre-test in September is also another limitation in the study. Because of the hectic and different schedules of each participant instructor, the researcher could not gather instructors in a classroom and force them to finish marking on the same day as in the pre-test. Therefore, not having the same assessment conditions might have led to differences between the two marking sessions.

Last but not least, because this study is solely based on quantitative data analysis, it could have been elaborated through qualitative tools like interviews or think-aloud protocol to better analyse the rationale behind the scores given by the instructors. Therefore, based on the limitations mentioned above, it would not be appropriate to generalize the findings of this study to other teachers at the same school or to those in other higher education institutions.

5.3 Implications for Testing and Practice

The current study has found some significant points for both testing and pedagogy that higher education institutions might take advantage of. The study has concluded that standardisation sessions are of utmost importance to reduce the possible rater inconsistencies to a great extent, though it does not eliminate them

altogether (Wood & Quinn, 1976). Assessment of direct writing methods are based on raters' subjective judgements and even a small discrepancy might affect test takers' exam results. Therefore, here are some recommendations that institutions may follow to eradicate rater inconsistency as much as possible.

To begin with, although rubric use (Jonsson & Svingby, 2007; Kohn, 2006; Moskal & Leyden, 2000 and Silvestri & Oescher, 2006) and double-marking protocol (Chaplen, 1969; Fearnley, 2005 and Wood & Quin, 1976) are effective in providing better consistency among raters, standardisation sessions, in which rubric and exam specifications as well as expectations from student performances are explained in detail, should be provided to the instructors by the institutions.

However, efficiency of these sessions is as important as their effectiveness on rater reliability (Pufpaff, Clarke & Jones, 2015). Due to their hectic schedule, teachers might not benefit from these sessions as much as expected. In other words, among all the duties they have at work, standardisation sessions might seem an extra burden so they might not concentrate on the sessions, which, as a result, might affect their grading. Therefore, institutions might look for ways to make these sessions more productive and less loaded for teachers. Maybe, instead of meeting face-to-face all the time, teachers might receive online sessions or they could discuss their ideas on a forum, which would not require them to be physically present at work.

Furthermore, it is of utmost importance to provide fairness in assessment. Therefore, any factors that might hinder this should be removed (Hughes, 2003). For instance, for a fairer grading, blind marking protocol, in which students' names are hidden from raters, should be followed. In this way, raters' potential subjective judgements for a test taker could be avoided.

5.4 Recommendations for Future Research

The current study suggests that standardisation sessions are required for a more reliable and consistent evaluation for direct written assessments. Therefore, based on the findings and the limitations mentioned in previous sections, there are some suggestions for future research to provide better evidence for the findings obtained in the study.

First of all, the study could be replicated in a different institution or country, with different participants to be able to better generalize the findings in the current study. Secondly, the same study with the same methodology could be replicated with more participants. To better investigate the effect of rater differences, a future study like native vs. non-native or lay vs. experienced raters, should especially include more of these teachers.

Moreover, each component of writing like language use, content or organization could be investigated in detail in the future. Unlike the current study which only focused on the total scores of the students, raters' scores awarded to each band could be studied individually for a deeper understanding of their assessment. Also, with a change in the methodology, the study can be re-enacted with a treatment group that has never attended such sessions before and a control group that continuously receives standardisation sessions before each written assessment. In this way, whether and to what extent not being aware of such practices affects raters' judgements could be better identified.

Furthermore, this quantitative study could be extended to a mixed one which also includes think-aloud protocol and interviews with the raters to be able to gain more insights as to why and how teachers give such scores in their two different

scorings. In addition, apart from studying face-to-face standardisation sessions, the effectiveness and efficiency of other types of standardisation meetings, like online sessions, could also be studied in a future study. Finally, the current study and the recommendations given above could also be applied to speaking exam performance, which is another direct assessment type that requires subjective rater judgement.

REFERENCES

Anadol, H. Ö. & Doğan, C. D. (2018). Dereceli puanlama anahtarlarının güvenirliğinin farklı deneyim yıllarına sahip puanlayıcıların kullanıldığı durumlarda incelenmesi. *İlköğretim Online*; *17(2),* 1066-1076. Available at: http://ilkogretim-online.org.tr

American Heritage® Dictionary of the English Language. 5th Edition*. (*2011). Retrieved December 31 2018 from https://www.thefreedictionary.com/validity

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

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

Baird, J. A., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education*, *11 (3)*, 331 - 348. https://doi.org/10.1080/0969594042000304627

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

Black, P. (1998). *Testing: Friend or foe?* London: Falmer Press.

Bukta, K. (2013). *Rating EFL written performance*. Chapter 4. London: Versita Ltd.

Brooks, V. (2004) *Double Marking Revisited*. British Journal of Educational Studies, *52 (1),* 29 - 46. Taylor & Francis, Ltd.

Brown, A. (1995). The Effect of Rater Variables in the Development of an Occupation-specific Language Performance Test. *Language Testing*, *12(1),* 1–15.

Brown, R. (1978). What we know and how we could know more about writing ability in America. *Journal of Basic Writing 1(4),* 1-6.

Brown, H.D. (2003). *Language assessment: Principles and classroom practice.* New York, NY: Longman.

Carrell, P.L. & Connor, U. (1991). Reading and writing descriptive and persuasive texts. *The Modern Language Journal*, *75,* 314-324.

Chaplen, E. F. (1969) The reliability of the essay subtest in a university entrance test in English for non-native speakers of English. In G. E. Perren, & J.L.M. Trim (Eds.), *Applications of Linguistics - papers from the second International Congress of Applied Linguistics, Cambridge, 1969.* Cambridge: Cambridge University Press.

Charney, D. (1984). The Validity of using holistic scoring to evaluate writing: A critical overview author(s). *Research in the Teaching of English*, *18(1),* 65-81. National Council of Teachers of English Stable. https://www.jstor.org/stable/40170979

Chapman, V. G., & Inman, M. D. (2009). A conundrum: rubrics or creativity/metacognitive development? *Educational Horizons*, *87(3),* 198-202.

Coffman, William E. 1971. On the reliability of ratings of essay examinations in English. *Research in the Teaching of English*, 5, 24–36.

Cooper, C. R. (1977). Holistic evaluation of writing. In Charles R. Cooper & Lee Odell (Eds.), *Evaluating writing: Describing, measuring, judging*, 3–31. Urbana: National Council of Teachers of English.

Davis, L. (2018). Primary trait scoring. In B. Frey (Ed.). *The SAGE encyclopedia of educational research, measurement, and evaluation*. 1296-1297. Thousand Oaks, CA: SAGE Publications, Inc.

Dawson, C. M. (2009). Beyond checklists and rubrics: engaging students in authentic conversations about their writing. *English Journal*, 98(5), 66-71.

Dempsey, M. S., Pytlikzillig L. M. & Burning, R.H. (2009). Helping service teachers learn to assess writing: Practice and feedback in a web-based environment. *Assessing Writing,* 14(1), 38-61.

Diederich, P. B., French J. W., & Carlton, S. T. (1961). Factors in judgement of writing ability. *Research Bulletin RB,* 61-15. Princeton, NJ: Educational Testing Service.

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

Dobrić, N. (2018). Reliability, validity, and writing assessment: A timeline. *ELOPE: English Language Overseas Perspectives and Enquiries*, *15*(2), 9-24. https://doi.org/10.4312/elope.15.2.9-24

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

Elder, C., Barkhuizen, G., Knoch, U., & Randow, J. V. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37-64.

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, 1(1), 1-16. Retrieved from https://www.researchgate.net/publication/228516924_The_Effects_of_Rater_Training_on_Raters27_Severity_and_Bias_in_Second_Language_Writing_Assessment

Fearnley, A. (2005) *An investigation of targeted double marking for GCSE and GCE*. London: Qualifications and Curriculum Authority. Available at: http://dera.ioe.ac.uk/9450/1/QCDA104979_an_investigation_of_targeted_double_marking_for_GCSE_and_GCE.pdf (accessed 26th March 2019).

Ghanbari, B., Barati, H. & Moinzadeh, A. (2012). Rating scales revisited: EFL writing assessment context of Iran under scrutiny. *Language Testing in Asia*, *2(1),*83–100.

Godshalk, F. I., Swineford, F. & Coffman, W. E. (1966). The measurement of writing ability. New York: College Entrance Examination Board.

Gonzalez, E.F. & Roux, R. (2013). Exploring the variability of Mexican EFL teachers' ratings of high school students' writing ability. *Argentinian Journal of Applied Linguistics, 1*(2), 61-78.

Gonzales, E.F., Trejo, N.P. & Roux, R. (2017). Assessing EFL University Students' Writing: A study of score reliability. *Revista Electrònica de Investigaciòn Educativa*, (19)2, 91-103. https://doi.org/10.24320/redie.2017.19.2.928

Greatorex J., Baird, J. & Bell, J. F. (2002, August). *Tools for the trade: what makes GCSE marking reliable?* Paper presented at the EARLI Special Interest Group on Assessment and Evaluation, University of Northumbria, UK, August.

Greatorex J., Shannon M. (2003, September). *How can NVQ assessors' judgements be standardised?* Paper presented at the British Educational Research Association Conference, Heriot-Watt University, Edinburg.

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

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

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing: An International Journal, (1),* 5-16.

Hamp-Lyons, L. & Davies, A. (2008). The Englishes of English test: bias revisited. *World English*, *27 (1),* 26-39.

He, T.-H., Gou, W. J., Chien, Y. C., Jenny Chen, I. S., & Chang, S. M. (2013). Multi-Faceted Rasch Measurement and Bias Patterns in EFL Writing Performance Assessment. *Psychological Reports, 112(2),* 469–485. https://doi.org/10.2466/03.11.PR0.112.2.469-485

Head, J.J. (1966) Multiple marking of an essay item in experimental O-level Nuffield Biology examinations. *Educational Review, 19 (1),* 65–71.

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

Hitt, A. M., & Helms, E. C. (2009). Best in show: Teaching old dogs to use new rubrics. *Professional Educator*, *33(1),* 1-15.

Huddleston, E. (1952*). Measurement of writing ability at the college-entrance level: Objective vs. subjective testing techniques*. (Doctoral dissertation). New York University, Princeton, New Jersey.

Hughes. A. (2003). *Testing for language teachers*. 2nd edition. Cambridge: Cambridge University Press.

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

Kayapınar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research*, 57, 113-136. http://dx.doi.org/10.14689/ejer.2014.57.2

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

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale.* Frankfurt: Peter Lang GmbH.

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

Kohn, A. (2006). The trouble with rubrics. *English Journal*, *95(4),* 12-15.

Kondo, Y. (2010). Examination of rater training effect and rater eligibility in L2 performance assessment. *Journal of Pan-Pacific Association of Applied Linguistics, 14(2),* 1-23.

Laming, D. (2004). *Human judgement: The eye of the beholder.* London: Thomson.

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

Lloyd-Jones. R. (1977). Primary-Trait Scoring. In Charles R. Cooper and Lee Odell (Eds.). *Evaluating writing: describing, measuring, judging*. (33–76). Urbana: National Council of Teachers of English.

Lovorn, M. Rezaei & A. R. (2011). Assessing the assessment: Rubrics training for pre-service and new in-service teachers. *Practical Assessment, Research & Evaluation, 16,* 1-18.

Lucas, A. M. (1971) Multiple marking of a matriculation biology essay question. *British Journal of Educational Psychology, 41 (1),* 78–84.

Lumley, T. (2006). *Assessing second language testing: the rater's perspective.* Frankfurt, Germany: Peter Lang.

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

Lunsford, A. (1986). The past - and future- of writing assessment. In K. L. Greenberg, H. S.Wiener, & R. A. Donovan (Eds.). *Writing assessment: Issues and strategies* (1–12). White Plains, NY: Longman.

Lunz, M. E., Wright, B.D. & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3(4),* 331-345.

Luo, K. K. (2015). Validity considerations in designing a writing test. *Studies in Literature and Language, 10 (5),* 19-21. Available from: http://www.cscanada.net/index.php/sll/article/view/6957

Maxwell, S. (2010). Using rubrics to support graded assessment in a competency based environment. Occasional paper. Australian Government Department of Education, National Centre for Vocational Education Research. Retrieved from https://files.eric.ed.gov/fulltext/ED509189.pdf

Meadows, M. & Billington, L. (2005) *A Review of the Literature on Marking Reliability.* National Assessment Agency. Retrieved from https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf

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

McColly, W. (1970). What does educational research say about the judging of writing ability? *Journal of Educational Research, 64(*4), 147-156.

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

Noyes, Edward S., William M. Sale, & John M. Stalnaker. (1945). *Report on the First Six Tests in English Composition*. New York: College Entrance Examination Board.

Ofqual (2014). *Review of double marking research.* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605661/2014-02-14-review-of-double-marking-research.pdf

Odell, L. (1981). Defining and assessing competence in writing. In Charles Cooper (Ed.). *The Nature and Measurement of Competency in English.* (95-136). Urbana, IL: National Council of Teachers of English.

Özel, S. & Acar, T. (2014, June). *Okullarda sınıf içi ölçmelerde G katsayısı*. Paper presented at IV. National Congress on Measurement and Evaluation in Education and Psychology, Hacettepe University, Turkey.

Parlak, B. & Doğan, N. (2014). Dereceli puanlama anahtarı ve puanlama anahtarından elde edilen puanların uyum düzeyleri. *Hacettepe University,*

Education Faculty Journal, *29(2),* 189-197.

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.

Picturesque Expressions: A Thematic Dictionary. 1st Edition. (1980). Retrieved December 31 2018from https://www.thefreedictionary.com/validity

Pilliner, A. E. G. (1969). Multiple marking: Wiseman or Cox? *British Journal of Educational Psychology*, *39(3),* 313-315.

Pufpaff, L.A., Clarke, L. & Jones, R.E. (2015). The Effects of Rater Training on Inter-Rater Agreement. *Mid-Western Educational Researcher, 27(2),* 117-141.

Quellmalz, E.S., Capell, El, & Chou, C. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement 19(4),* 241-258.

Raikes N., Fidler J. & Gill T. (2009, September). *Must examiners meet in order to standardise their marking? An experiment with new and experienced examiners of GCE AS Psychology.* Paper presented at the British Educational Research Association Annual Conference, University of Cambridge, UK.

Random House Kernerman Webster's College Dictionary. (2010). Retrieved December 31 2018 from https://www.thefreedictionary.com/validity

Reddy, Y. M., & Andrade, H. L. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35(4),* 435-448.

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

Robinson, D. (2000). Building consensus on the scoring of students' writing: A comparison of teacher scores versus native informants' scores. *The French Review, 73*(4), 667-688. Retrieved from http://www.jstor.org/stable/398603

Sadler, D. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18(2),* 119-144.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing*, *14(2),* 157-184. https://doi.org/10.1177/026553229701400203

Shaw, S.D. (2002). IELTS writing: Revising assessment criteria and scales (phase 2). *Research Notes. 10(4),* 1-24. Retrieved on December 17, 2018 from http://www.cambridgeesol.org/rs_notes/rs_nts10.pdf

Shaw, S.D. & Weir, C.J. (2007), *Examining writing: research and practice in assessing second language writing.* Cambridge: Cambridge University Press.

Shavelson, R.J., Gao, X., & Baxter, G. (1996). On the content validity of performance assessments: centrality of domain-specifications. In M. Birenbaum & F. Dochy (Eds.). *Alternatives in assessment of achievements, learning process and prior knowledge.* Boston: Kluwer Academic Publishers.

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

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18(3),* 303–325. https://doi.org/10.1177/026553220101800303

Shi, L., Wan, W. & Wen, Q. (2003). Teaching experience and evaluation of second language students' writing. *The Canadian Journal of Applied Linguistics, 6*, 219-236.

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.

Silvestri, L. & Oescher, J. (2006). Using rubrics to increase the reliability of assessment in health classes. *International Electronic Journal of Health Education, 9,* 25-30.

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

Stiggins, R. J. (1982). A comparison of direct and indirect writing assessment methods. *Research in the Teaching of English*, *16(2),* 101-114. Retrieved from http://www.jstor.org/stable/40170937

Sweedler-Brown, C.O. (1993). ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing*, *2(1),* 3-17.

Tajeddin, Z., Alemi, M., & Pashmforoosh, R. (2011). Non-native teachers' rating criteria for L2 speaking: Does a rater training program make a difference? *Teaching English Language*, *5(1),* 125-153.

Tanrıverdi-Köksal, F. (2013). *The effect of raters' prior knowledge of students' proficiency levels on their assessment during oral interviews.* (Unpublished master's thesis). Bilkent University, Ankara.

Tisi J., Whitehouse G., Maughan S. & Burdett, N. (2013). *A review of literature on marking reliability research* (report for Ofqual). Slough, NFER. Available at: www.ofqual.gov.uk/files/2013-06-07-nfer-a-review-of-literature-on-marking-reliability.pdf

Venkatasamy, V.E. (2016). Influence of rater's experience and background: Implications on raters training. *ELT Vibes: International E-Journal for Research in ELT, 2 (4),* 44-53.

Wang, B. (2010). On rater agreement and rater training. *English Language Teaching*, *3(1),* 108-112.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15(2),* 263-287.

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

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10,* 305-335.

Wiseman, S. (1949) The marking of English composition in grammar school selection. *British Journal of Educational Psychology*, *19 (3),* 200–209.

Wood, R. & Quinn, B. (1976) Double impression marking of English language essay and summary questions. *Educational Review*, *28(3),* 229-246.

APPENDICES

A. INFORMED CONSENT FORM

Dear Colleague,

I am M. Nur Karadenizli Çilingir. I have been working as an English Instructor in English Language School at TED University since September 2014. I am doing my Master's degree in English Language Teaching at Middle East Technical University. The subject of my thesis study is to investigate whether and to what extent the standardisation meetings conducted before writing exams improve and contribute to inter- and intra-rater reliability.

In this study, the related data gathered from the participants will be acquired through two different marking sessions once with and once without standardisation meetings. In the first session, the participant teachers will be given a standardisation session before they assess the students' exam papers. 3-4 months later, the second session will be conducted without a standardisation session. Therefore, you are required to attend both of the marking sessions. The scorings gathered from two different marking sessions will be analysed quantitatively. The data provided in the study will be used for scientific research purposes only.

Participation in the study is on a voluntary basis. There are no known risks associated with this study, however, provided that either you feel uncomfortable or do not want to continue the study, please feel free to withdraw from the study at any time. In the event you choose to withdraw from the study, all information you have provided will be destroyed and omitted from the final paper.

The data gathered in the study will be treated confidentially and will not be linked to any individual names or addresses.

Your participation will contribute to the study to a great extent. If you accept taking part in this study, please fill in the related blanks at the bottom of this page and sign.

If you want to learn more about the details or be informed about the findings of the study, please contact with me via the following e-mail address.

Thank you for your invaluable contribution.

M. Nur KARADENİZLİ ÇİLİNGİR nur.cilingir@metu.edu.tr

Advisor: Prof. Dr. Gölge SEFEROĞLU

MA, Middle East Technical University / ANKARA

By signing this consent form, I certify that I agree to the terms of this agreement.

Name and Surname:	Signature:	Date:
…… ……	…….	…../…../…..

B. DEMOGRAFIC INFORMATION QUESTIONNAIRE

Dear Colleague,

This questionnaire was designed to obtain some background information about the teachers participating in this thesis study. All of your answers will be kept confidential and they will not be associated with your name.

Please answer all the questions provided below.

1) Age: ……...

2) Gender: ☐ Male ☐ Female

3) Nationality: ☐

4) Graduated B.A Program:

a) English Language Teaching

b) English Language and Literature / American Culture and Literature

c) Translation and Interpretation

d) Linguistics

e) Other ……

5) MA degree: a) No b) Yes, continuing c) Yes, completed

If yes, please specify your field:

a) English Language Teaching

b) English Language and Literature / American Culture and Literature

d) Linguistics

e) Educational Sciences

f) Other: …….

6) **PhD:** a) No b) Yes, continuing c) Yes, completed

If yes, please specify your field:

a) ELT

b) English Language and Literature / American Culture and Literature

c) Educational Sciences

d) Other: ……

7) Years of experience in teaching English: ……

8) How long have you been working at this Institution?:..... years

9) How long have you been marking writing exams?:..... years

10) Have you taken any courses or attended any training programs and/or workshops related to rater training (except for the standardisation sessions conducted before exams at your current university or any other university you worked at before)?

a) Yes b) No

If yes, please specify:

1) ……
2) ……
3) ……
4) ……
5) ……

Thank you for your invaluable time and contribution.

M. Nur KARADENİZLİ ÇİLİNGİR nur.cilingir@metu.edu.tr
Advisor: Prof. Dr. Gölge SEFEROĞLU
MA, Middle East Technical University / ANKARA

C. APPROVAL OF THE METU HUMAN SUBJECTS ETHICS COMMITTEE

UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59
ueam@metu.edu.tr
www.ueam.metu.edu.tr

Sayı: 28620816 / 198

09 Nisan 2019

Konu: Değerlendirme Sonucu

Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi: İnsan Araştırmaları Etik Kurulu Başvurusu

Sayın Prof.Dr. Gölge SEFEROĞLU

Danışmanlığını yaptığınız **Mahmure Nur Karadenizli ÇİLİNGİR'in "The Effect of Standardisation Meetings Conducted Before English Language Writing Exams on the Improvement of Inter-Reliability and Intra-Rater Reliability"** başlıklı araştırması İnsan Araştırmaları Etik Kurulu tarafından uygun görülmüş ve **185-ODTÜ-2019** protokol numarası ile onaylanmıştır.

Saygılarımızla bilgilerinize sunarız

Prof. Dr. Tülin GENÇÖZ
Başkan

Prof. Dr. Ayhan SOL
Üye

Prof. Dr. Ayhan Gürbüz DEMİR
Üye

Prof. Dr. Yaşar KONDAKÇI
Üye

Doç. Dr. Emre SELÇUK
Üye

Doç. Dr. Pınar KAYGAN
Üye

Dr. Öğr. Üyesi Ali Emre TURGUT
Üye

D. TURKISH SUMMARY / TÜKÇE ÖZET

İngilizcenin ortak dil olduğu modern dünyada, öğrencilerin dili ne kadar öğrendiklerini doğru bir şekilde ölçebilmek çok önemlidir. Bu nedenle, son zamanlarda, öğrencilerin dil kazanımını doğru bir şekilde değerlendirmek yükseköğretim kurumlarının oldukça önem verdikleri bir alandır. Dolayısıyla, öğrencilerin dil gelişimini değerlendirmek için birçok farklı prosedür ortaya çıkmıştır. Yazma becerisi sınavları bunlardan bir tanesidir. Mevcut yazma becerisi değerlendirme yöntemleri *dolaylı* ve *doğrudan* olarak ikiye ayrılmaktadır. Charney bu iki metodu sırasıyla *nicel* ve *nitel* olarak da adlandırmaktadır (1984). Bunun nedeni, dolaylı yazma becerisi sınavlarının değerlendirmesi nesnel iken, doğrudan yazma becerisi sınavlarının değerlendirmesinin puanlayıcıların öznel yorumuna ve değerlendirmesine dayalı olmasıdır.

Dolaylı yazı becerisi sınavlarında genellikle öğrencilerden bir dizi alternatiften doğru cevabı seçmeleri ya da birbirleriyle uyuşan cümle ya da sözcük öbeklerini eşleştirmeleri istenir. Bu değerlendirme metotu genellikle çoktan seçmeli, doğru / yanlış veya eşleştirme soru türlerinden oluşmakta ve özellikle dilbilgisi, imla, noktalama, cümle kurma ve benzeri yazı kurallarını ölçmeyi amaçlamaktadır (Stiggins, 1982). Yani, bu değerlendirme türü, öğrencilerin yazma yetilerini direk ölçmek yerine, yazma becerisini oluşturan ögeler hakkında ne kadar bilgi sahibi olduklarını test etmektedir. Dolaylı yazma becerisi değerlendirme yöntemlerini tatbik etmek genellikle kolaydır. Ayrıca, öğrencilerin dil yeterliliğini kısa sürede değerlendirebilmek adına oldukça pratiktirler. Bu tür sınavları notlandıran kişiler,

değerlendirme boyunca öznel yargılarını kullanmazlar. Sınav kağıtları önceden hazırlanmış cevap anahtarıyla öğretmenler ya da optik okuyucu tarafından notlandırılr. Dolayısıyla, bu tür sınavların puanlaması nesnel bir değerlendirmeye dayandığı için yüksek güvenilirliğe sahiptir.

Güvenilirlik, öğrencilerin performansını ölçmek için değerlendirme aracının tutarlılığı ve doğruluğu olarak tanımlanmaktadır (Bachman & Palmer, 1996). Yani, güvenilir bir yazma becerisi değerlendirmesi, farklı öğrencilere veya farklı zamanlarda aynı öğrencilere uygulandığında ya da farklı puanlayıcılar tarafından değerlendirildiğinde benzer sonuçlar vermelidir (Stiggins, 1982). Güvenirlik puanlayıcı (inter-rater) ve puanlama (intra-rater) güvenirliği olmak üzere ikiye ayrılır. Puanlayıcı güvenirliği, aynı sınavı değerlendiren iki ya da daha fazla kişinin değerlendirmelerinin arasındaki tutarlılıkken, puanlama güvenirliği, bir puanlayıcının farklı zamanlardaki değerlendirmeleri arasındaki tutarlılığa verilen isimdir.

Ölçme ve değerlendirme alanındaki birçok önemli isim, bir değerlendirme yönteminde yüksek güvenirlik elde edebilmek için, o ölçeğin öznellikten arındırılmış olması gerektiğini savunmaktadır. Noyes, Sale ve Stalnaker (1945), Diederich (1950) Huddleston (1952) ve Godshalk, Swineford ve Coffman (1966) yüksek güvenilirlik sunabilecek değerlendirme yönteminin sadece dolaylı değerlendirme metotları olduğunu savunmaktadırlar.

Öte yandan, alandaki birçok önemli araştırmacı, öğrencilerin yazma yetisini ölçebilmenin tek yolunun, onlara ‘yazı yazdırmak ‘olduğunu, bunun da ancak doğrudan değerlendirme yöntemleriyle mümkün olabileceğini ifade etmişlerdir. Dolayısıyla, doğrudan değerlendirme yöntemlerini savunanlar, güvenirlik ile birlikte

ölçme ve değerlendirme alanındaki diğer önemli unsuru gündeme getirmişlerdir: *geçerlik*.

Geçerliğin birçok sözlük tanımı vardır: 'Geçerli olma durumu veya niteliği'(Random House Kernerman Webster's College Dictionary, 2010); 'Geçerli, sağlam ve savunulabilir olmak; teste alındığında tutarsızlık veya eksiklik göstermemek' (Picturesque Expressions: A Thematic Dictionary, 1980); 'Mantıksal kusurlardan uzak durmak ya da geçerli akıl yürütmeye dayanmak' (American Heritage Dictionary of the English Language, Fifth Edition, 2016). Ölçme değerlendirme açısından ise geçerlik, bir değerlendirme yönteminin doğruluğunu ve bu yöntemin ölçmesi gerekeni ölçüp ölçmediğini ifade eder. Yani bir testin ölçmek istediği şeyi doğru bir şekilde ölçmesi o testin geçerli olduğu anlamına gelir (Hughes, 2003, s.22). Dolaylı yazma beceri yöntemlerine karşı olanlar, öğrenciden *yazmasını* istemediği, yani ölçmesi gereken yetiyi ölçmediği için, bu değerlendirme türlerinin geçersiz olduğunu savunmuşlardır (Cooper, 1977 ve Luo, 2015). Kısacası, bir test ancak ölçmeyi hedeflediği şeyi ölçerse geçerli olabileceği için, dolaylı yazma becerisi yöntemlerinin geçerli olmadığını, örneğin, çoktan seçmeli sınav türlerinin, öğrencilere birçok yanlış alternatifin içinden doğru cevabı seçtirmenin ötesine geçemeyeceğini öne sürmüşlerdir. Brown, çoktan seçmeli soruların pasif bir zihinsel durum gerektirdiğini, ancak doğrudan yazma becerisi yöntemlerinin, öğrencilere organizasyon veya mekanik gibi yazma becerisi ögelerini kullandırtarak, zihinlerinin aktif kalmasını sağladığını savunmuştur (1978).

1970'lerin son zamanlarının ve 1980'lerin odak noktası, yazılı değerlendirmedeki geçerlik meselesiydi. O zamana kadar, birçok kişi çoktan seçmeli testlerin ya da diğer dolaylı değerlendirme yöntemlerinin yazma becerisine gereken

önemi vermediğinden ve öğrencileri noktalama, heceleme ya da dilin kullanımı gibi dilin unsurlarını ayrı ayrı ezberlemeye ittiğinden şikayet etmekteydi (Witte ve diğerleri, 1986). Bachman'ın özellikle geçerlik konusundaki argümanları bu seslerin daha fazla duyulmasına yardımcı oldu. Örneğin, dolaylı değerlendirme yöntemlerinin içerik geçerliğine sahip olmadığı gündeme getirilmiştir. Dolaylı ölçme metotları öğrencilerin, fikirlerini bir araya getirme, onları tekrar gözden geçirme ve düzenleme gibi yazma yetisinin önemli unsurlarını kullanmalarına izin vermiyordu. (Brown, 1978 ve Cooper, 1977). Ayrıca, bu tür sınavların yapı geçerliği olmadığı da belirtilmiştir. Yani, öğrenciler yazmak yerine bir dizi yanlış alternatif arasından doğru cevabı bularak ya da eşleştirme sorularına cevap vererek değerlendirilmekteydi. Brown'un da belirtildiği gibi, öğrencilerin yazma yetileri, tek bir kelime bile yazmalarını gerektirmeden ölçülmek istenmekteydi (1978). Geçerlik, bir testin söz konusu davranışı ölçüp ölçemediği ve ne kadar iyi ölçebildiği anlamına geldiğine göre; dolaylı değerlendirme karşıtları, öğrencilerin yazma yetilerinin, onlara yazı yazdırmadan ölçmenin imkansız olduğunu, dolayısıyla bu değerlendirme türlerinin geçerliği olmadığını ileri sürmüşlerdir.

Ancak özellikle 1970'lerde doğrudan değerlendirme yöntemlerinin popülerlik kazanması, bu değerlendirme metodunun güvenirliği ile ilgili tartışmaları ve endişeleri alevlendirmiştir. Örneğin, alandaki birçok önemli isim doğrudan değerlendirme yöntemlerinin güvenirliğinin olmadığını, dolayısıyla dolaylı ölçme yöntemlerinin kullanılmaması gerektiğini savunmuştur (Huddleston 1952; Sheppard 1929; Stalnaker, 1936; Starch & Elliot 1912 ve Traver & Anderson 1935). Özellikle Diederich, French ve Carlton'ın 1961 yılındaki kapsamlı çalışması, dolaylı ölçme yöntemlerinin güvenirliğinin olmadığı görüşünü tasdiklemiştir. 300 üniversite

öğrencisinin yazma becerisi sınav kağıtları, İngilizce öğretmeni, Sosyal Bilimler Öğretmeni, Fen Bilimleri Öğretmeni, yazar - editör ve avukat - iş adamlarından oluşan beş farklı meslek grubuna ait toplamda 53 puanlayıcı ile değerlendirilmiştir. Çalışmanın sonunda, puanlayıcıların öğrenci kağıtlarını (1) fikirler, (2) biçim, (3) mekanik, (4) ifade ve (5) stil olmak üzere beş farklı kategoriye ayırarak değerlendirdikleri ortaya çıkmıştır. Bazı puanlayıcılar fikir ve içerik üzerinde dururken, bazılarınınsa dil kullanımına ya da mekaniğe önem verdiği gözlenmiştir. Dolayısıyla, çalışmayı yapan araştırmacılar, doğrudan yazma becerisi metotlarının öğrencilerin dil yeterliliğini değerlendirmek için güvenilir bir yöntem olarak kullanılamayacağını, çünkü her bir puanlayıcının, değerlendirme sırasında öznel yargılarını kullandığı sonucuna varmıştır. Ancak, ölçme ve değerlendirme alanındaki bazı kişilere göre, çalışmada bazı eksiklikler vardı (Braddock, Llyod-Jones & Schoer, 1963 ve Dobrić, 2018). Örneğin, puanlamayı yapan kişilere rubrik ve değerlendirme sırasında dikkat etmeleri gereken hususlar hakkında bilgilendirme yapılmaması bunlardan bazılarıdır. Değerlendirmeyi yapan kişiler, tüm kararları kendi başlarına vermek zorunda kalmışlardır.

1970'ler, güvenirliğinin yüksek olması nedeniyle dolaylı değerlendirme yöntemlerinin kullanılması gerektiğini savunan ve bu yöntemlerin yetersizliği konusunda hemfikir olan birçok araştırmacıya tanık olmuştur. Dolaylı değerlendirme metotlarına karşı olanlar, özellikle sınıf içindeki yazma becerisi uygulamaları ile sınav sırasında öğrenciden beklenenin birbiriyle örtüşmediğini öne sürmüştür. Dolaylı değerlendirme yöntemlerinin 'insanların tek bir kelime yazmalarını istemeksizin yazma yeteneklerini test ettiğini' (Brown, 1978, s.1), dolayısıyla bunun kabul edilemez olduğunu savunmaktadırlar.

Doğrudan değerlendirme metotlarının güvenilir olup olmadığı tartışması zamanla yerini öznel ölçme metotlarının nasıl daha güvenilir olabileceği arayışına bırakmıştır. Örneğin Coffman puanlamada farklılıkların olmasının nedeninin puanlamayı yapan kişilerin yargılarındaki farklılıklardan veya aynı puanlayıcının kararlarının zaman içinde değişkenlik göstermesinden kaynaklanabileceğini; dolayısıyla puanlamayı yapacak kişilere testin güvenirliğini düşürecek etmenlere ilişkin farkındalık kazandırmak gerektiğini ileri sürmüştür (1977, s.36).

Coffman gibi, ölçme ve değerlendirme alandaki diğer birçok isim, güvenilirliği etkileyebilecek olası nedenlere odaklanmıştır. Örneğin, bir değerlendirme yönteminin güvenirliğini düşüren etmenlerden bir tanesinin testi değerlendirenlerin belirli ön yargıları olduğunu öne sürmüştür (Scheafer, 2008; Schoonen, Vergeer & Eiting, 1997 ve Sherwin, 1969). Benzer şeklide Shi çalışmasında, anadilin de puanlayıcıların yargılarını etkilediği sonucuna varmıştır (2001). Ana dili Çince ve İngilizce olan 46 deneyimli öğretmenin İngilizce öğrenci makalelerini değerlendirdiği çalışması, iki öğretmen grubunun notlarının benzerlik gösterdiğini, ancak ana dili Çince olan öğretmenlerin ana dili İngilizce olan öğretmenlere kıyasla organizasyon ve içeriğe daha fazla önem verdiği ve bu unsurları daha katı bir şekilde puanladığı sonucuna ulaşmıştır. Çalışma, iki farklı öğretmen grubunun iyi bir yazma becerisine sahip bir öğrencinin taşıması gereken unsurlar konusunda farklı bir anlayışa sahip olduğunu; bu nedenle, öğretmenlerin ana dillerinin puanlamalarını etkileyebilecek bir faktör olabileceğini gündeme getirmiştir.

Bir değerlendirme yönetiminin güvenilirliğini olumsuz etkileyecek bir diğer olası faktör de puanlamayı yapan kişilerin puanlama deneyimidir. Alandaki çalışmalar çoğunlukla, puanlayıcılar ne kadar az deneyimli olursa, puanlayıcı güvenilirliğinin o kadar düşük olduğunu savunmaktadır. Örneğin, Barkaoui (2007), puanlama deneyimi olmayan öğretmenlerin deneyimli öğretmenlere nazaran rubriği yorumlamada zorluk çektiklerini, dolayısıyla, puanlamanın güvenirliğini tehlikeye soktuklarını ileri sürmüştür. Hatta bazı tecrübeli öğretmenlerin bile rubriği farklı şekilde yorumlayabildiği öne sürülmüştür (Brown, 2003; Hamp-Lyons & Davies, 2008; He, Gou, Chien, Chen & Chang, 2013; Lumley, 2006 ve McNamara, 1996). Kısacası literatürde, doğrudan değerlendirme yöntemlerinin güvenilir bir şekilde değerlendirilmesinin çok kolay olmadığını, çünkü değerlendirmeyi yapan kişilerin yargılarını etkileyebilecek birçok faktörün olduğunu savunan birçok çalışma vardır.

Bu amaçla, 20. yüzyılın ilk yarısında puanlayıcıların tutarsızlığını ortadan kaldırabilmek - ya da azaltabilmek - ve yazma becerisi yöntemlerinin güvenilirliğini artırmak adına birçok girişimde bulunulmuştur (Brooks, 2004). Bunlardan bir tanesi değerlendirme sırasında bir dereceli puanlama anahtarı, yani rubrik, kullanmaktır. Rubrik öğrencinin yazma çalışması gibi nitel bir değerlendirme gerektiren ölçme tekniklerinin değerlendirilmesinde kullanılan bir puanlama aracıdır (Jonsson & Svingby, 2007, s. 131). Birçok farklı rubrik türü vardır, ancak özellikle Bütüncül, Analitik ve Temel Özelliklere Dayalı Dereceli Rubrik türleri yaygın olarak kullanılmaktadır. *Bütüncül rubrik*, öğrenci performansını bütünsel olarak değerlendirir. Pratiktir, çünkü genel performansı değerlendirmeyi hedeflediği için değerlendirmeyi yapan kişilere zaman kazandırır. *Analitik rubrik* ile öğrencilerin performansları, içerik, dil kullanımı, dilbilgisi veya mekanik gibi farklı dil

unsurlarına ayrılarak detaylı bir şekilde değerlendirilir. *Temel Özelliklere Dayalı Dereceli Rubrik* ise öğrenci performansını daha çok öğrencilerin belirlenen amaçları yerine getirebilme düzeylerini açısından değerlendirir (Lloyd-Jones, 1977).

Birçok çalışma, rubrik kullanımının değerlendirme güvenirliğini arttırdığını öne sürmüştür. Örneğin, Anadol ve Doğan çalışmalarına katılan öğretmenleri üç grupta toplamıştır (2018). İlk grup, rubrik kullanımında bir yıldan az deneyime sahip öğretmenlerden oluşurken ikinci grup beş yıldan fazla deneyime sahip olanlardan ve son grupsa hem deneyimli hem de deneyimsiz öğretmenlerden oluşmaktadır. Çalışma, değerlendirme güvenirliğini arttırmada, rubrik kullanma deneyiminin değil, değerlendirme sırasında kullanılacak iyi tasarlanmış bir rubriğin etkili olduğunun sonucuna varmıştır. Benzer şekilde literatürde, rubriğin güvenirliğe katkıda bulunduğunu öne süren birçok isim vardır (Jonsson & Svingby, 2007; Kohn, 2006 ve Silvestri & Oescher, 2006). Moskal ve Leyden'a göre, iyi tasarlanmış bir rubrik, hem puanlayıcı hem de puanlama güvenirliğinin artmasını sağlar çünkü değerlendirme sırasında öznel yargının azalmasına yardımcı olur (2000).

Rubrik kullanımın yanında, nitel değerlendirmede güvenirliği arttırmak için yapılan girişimlerden bir diğeri de çoklu/ikili puanlandırma yöntemidir. Yani değerlendirme iki ya da daha fazla kişi tarafından yapılmaktadır. Genellikle öğrencinin notu, puanlayıcıların notlarının ortalaması alınarak hesaplanır. İlk defa 1941 yılında Wiseman tarafından uygulanan çoklu puanlama yönteminden, zamanla, daha pratik olması ve yine yüksek değerlendirme güvenirliği sağlaması nedeniyle ikili puanlama yöntemine geçilmiştir.

Ancak, gerek çoklu/ikili puanlama yönteminin gerekse rubrik kullanımının tek başına güvenilirliği arttırmayı garantilemediği ortaya atılmış ve birçok çalışmayla

desteklenmiştir. Alandaki birçok önemli isim, özellikle rubrik kullanımı üzerine eğilmiş; rubriğin değerlendirmeyi yapacak kişilerce en iyi şekilde özümsenmesi gerektiğini, bunun için de bu kişilere eğitimler verilmesi gerektiğini savunmuşlardır (Chapman & İnman, 2009; Dawson 2009; Gonzales, Trejo & Roux 2017; Kohn, 2006; Reddy & Andrade, 2010 ve Rezai & Lovorn, 2010).

Ölçme ve değerlendirme alanında bu eğitimlere verilen genel ad 'Puanlayıcı Eğitimi'dir. Puanlayıcı eğitimi bazen öğretmenlere düzenli bir şekilde verilen eğitimler silsilesiyken, bazen de her bir nitel değerlendirmeden önce, puanlayıcıların bir araya getirilip, gerek rubriğin gerekse sınavla ilgili detayların tartışıldığı ve puanlayıcıların ortak bir paydada buluşmasının sağlandığı kısa toplantılara verilen isimdir. Böylelikle bu toplantılar, puanlandırma sırasında nelere dikkat edilmesi gerektiğinin puanlayıcılarla tartışılmasını ve puanlayıcıların puanlama sırasında beklenmeyen durumlarla başa çıkmalarını sağlar (McNamara, 1996, s. 92). Bu toplantılara standardizasyon ya da kalibrasyon toplantıları da denmektedir. Çalışmanın yapıldığı üniversitede bu toplantılar standardizasyon toplantısı şekliyle adlandırıldığı için, bu çalışmada 'standardizasyon' terimi kullanılmaktadır.

Bu toplantılarda genellikle ilk olarak rubrik tartışılıp, değerlendirilmesi yapılacak sınav için geçerli detaylar Sınav Biriminden bir kişi tarafından puanlamayı yapacak öğretmenlerle paylaşılır. Sonrasında genellikle zayıf, orta ve iyi öğrenci kağıtlarının, öğrenci isimleri gizlenerek öğretmenler tarafından değerlendirilmesi istenir. Değerlendirme sonrasında, öğretmenler verdikleri puanları aralarında tartışarak, değerlendirme sırasında benzer kağıtlarla karşılaşıldığında nasıl bir tutum içinde olmaları gerektiğine dair ortak kararlar alırlar. Nitel ölçme metotlarının değerlendirilmesi puanlayıcıların öznel yargılarına dayalı olduğu için, bu

puanlamalar arasında bazı farklılıklar kabul edilebilir, ancak sürekli çok yüksek veya çok düşük puanlayanlar standartlarını düzenlemelidir (Knoch, 2009, s. 31). Dolaysıyla, standardizasyon toplantılarının amacı değerlendirmeyi yapacak kişilerin standartlarını düzenlemelerini sağlamaktadır. Kısacası, standardizasyon toplantıları, (1) değerlendirmeyi yapacak kişilerin her durumda geçerliğini koruyabilecek kararlar almasını; (2) tüm öğretmenlerin aynı bulguları temel alarak puanlama yapmasını ve (3) tüm öğrencilerin adil biçimde değerlendirilmesini sağlar (Greatorex & Shannon 2003, s. 3).

Literatürde, standardizasyon toplantılarının hem puanlayıcı (inter-rater) hem de aynı puanlama (intra-rater) güvenirliğini arttırdığını kanıtlayan birçok çalışma vardır (Brown, 1995; Elder, Barkhuizen, Gonzales, Trejo & Roux, 2017; Fahim & Bijani, 2011; Kayapınar, 2014; Knoch & Randow, 2007; Kondo, 2010; Lovorn & Rezai, 2011; Lumley & McNamara, 1995; Lunz, Wright & Linacre, 1990; Robinson, 2000; Shohamy, Gordon & Kraemer, 1992, Sweedler-Brown, 1985; Tajeddin, Alemi & Pashmforroosh, 2011; Wang, 2010 ve Weigle, 1999). McNamara, bu toplantıların asıl amacının farklı puanlayıcıların değerlendirmeleri arasındaki tutarsızlığı ortadan kaldırmaktan ziyade (puanlayıcı güvenirliği), aynı puanlayıcının kendi içinde tutarlı olmasını sağladığını (puanlama güvenirliği), çünkü puanlama güvenirliğinin sağlanması halinde zaten puanlayıcı güvenirliğine de ulaşılabileceğinin altını çizmektedir (1996).

Görüldüğü üzere, literatürde, değerlendirme güvenirliği üzerine birçok çalışma vardır. Ancak, hem puanlayıcı (inter-rater) hem de puanlama (intra-rater) güvenirliğini aynı anda ele alan çok az çalışma bulunmaktadır. Dolayısıyla bu çalışma, standardizasyon toplantılarının, hem puanlayıcı hem de puanlama

güvenirliği üzerine etkisini araştırması nedeniyle, literatürdeki birçok çalışmadan farklıdır.

Bu çalışma, Ankara'da bir vakıf üniversitenin hazırlık okulunda çalışmakta olan 24 İngilizce öğretim görevlisi ile yapılmıştır. Çalışma iki aşamadan oluşmaktadır. İlk aşamada, her bir katılımcı öğretim görevlisi, Eylül, 2018'de hazırlık atlama sınavının (EPE) yazma becerisi kısmına ait 22 öğrenci kağıdını değerlendirmiştir. Tüm hocalar, değerlendirme öncesinde bir standardizasyon toplantısına tabi tutulmuşlardır. Bu toplantıda, ilk önce sınav ile ilgili detaylar ve değerlendirmede kullanılacak analitik rubrik tartışılmış; daha sonrasındaysa sınava giren öğrenci kağıtları arasından Sınav Birimi tarafından rastgele seçilen beş öğrenci kağıdı, notlandırmayı yapacak öğretim görevlilerince bireysel olarak değerlendirilmiştir. Öğretim görevlileri, standardizasyon toplantısından hemen sonra kağıtları değerlendirmeye başlamıştır. Kendilerine verilen ilk grup öğrenciyi bitiren hocalar, ilk değerlendiren hocanın notunu görmemek şartıyla, başka bir grup öğrenciyi tekrar değerlendirmişlerdir. İki hocanın notları arasındaki fark en fazla dört ise, bu iki notun ortalaması öğrencinin hazırlık atlama yazma becerisi sınav notu olarak kaydedilmiştir. Ancak, iki hocanın değerlendirmesinin arasındaki fark beş ve üzeri olan kağıtlar, Sınav Birimi hocaları tarafından oluşturulan bir Komite tarafından üçüncü kez değerlendirilmiştir.

Çalışmaya katılan öğretim görevlilerinin Eylül ayında okudukları kağıtları hatırlamasını ve verdikleri notların etkisinde kalmalarını önleyebilmek adına, çalışmanın ikinci aşaması sekiz ay sonra, Mayıs, 2019'da tekrar edilmiştir. Çalışmanın ilk aşamasında her bir hoca tarafından okunan 22 adet öğrenci kağıdının içinden araştırmacı tarafından 10'ar adet kağıt rastgele seçilmiştir. Bu aşamada,

katılımcı hocalardan bir standardizasyon toplantısına katılmadan bu kağıtları tekrar değerlendirmeleri istenmiştir. Aynı ilk aşamada olduğu gibi, her bir öğrenci kağıdı iki hoca tarafından değerlendirilmiştir. Dolayısıyla, hocalar partnerler şeklinde kağıtları değerlendirmiş, böylece toplamda 12 partner oluşturmuşlardır. Her bir partner 10'ar öğrenci kağıdı; dolayısıyla, toplamda 120 öğrenci kağıdı değerlendirilmiştir. İlk aşamada da aynı öğrenci kağıtlarını değerlendiren hocalar, toplamda 240 adet kağıdı iki kez değerlendirmişlerdir. Çalışmanın ikinci aşamasında, birbirlerinden farklı ve yoğun programları olması sebebiyle, ilk aşamadan farklı olarak öğretim görevlilerinden kağıtları bir sınıfta toplanarak okumaları istenememiş; bunu yerine kendilerine kağıtları değerlendirmeleri için bir haftalık bir süre verilmiştir. Ancak sınavları puanlayıcıların kendi haline bırakılarak değerlendirmeleriyle ilgili soru sorabilecekleri bir platform oluşturulmaması, Diederich'in 1961 yılındaki çalışmasının eleştirilen kısımlarından biri idi. Bu nedenle, araştırmacı tarafından, katılımcı hocaların soru sorabilecekleri ve ihtiyaç duyduklarında gerekli yönlendirmeyi alabilecekleri bir *WhatsApp* grubu kurulmuştur.

Öğretim görevlilerinin birinci ve ikinci değerlendirmeleri IBM SPSS 23 yazılımı kullanılarak incelenmiştir. Sonrasında, parametrik ya da parametre dışı testlerden hangisinin uygulanacağının belirlenmesi için, ham veriye normallik testi uygulanmıştır. Buna göre iki bağımlı sayısal değişken (1.değerlendirme ve 2.değerlendirme) arasındaki farklılıklar Wilcoxon Signed Ranks Testi veri çözümlemesi ile incelenmiştir.

Wilcoxon Signed Test veri çözümlemesi sonuçlarına göre, sadece 10 öğretim görevlisinin ilk ve ikinci değerlendirmelerinin ortalamalarının arasında anlamlı bir fark görülmüştür. Fakat, öğretim görevlilerin ilk ve ikinci değerlendirmesi bireysel

incelendiğinde, tüm hocaların iki değerlendirmesi arasında farklılıklar olduğu saptanmıştır. Yazma becerisi sınavlarının nitel değerlendirme olması sebebiyle hocaların değerlendirmeleri arasında farklılıkların olması normal karşılansa da, hazırlık atlama sınavlarının öğrenciler için önemi ve onların eğitim hayatını ciddi şekilde etkileyebileceği göz önüne alındığında, bir puan bile öğrencilerin bölüme geçmelerini sağlayabilmekte ya da hazırlık eğitimini almak durumunda bırakabilmektedir. Dolayısıyla, bir puan çok ciddi bir fark gibi görünmese de, birçok öğrenci için çok büyük bir önem taşımaktadır. İki kez okunan toplam 240 öğrenci kağıdından 210 kağıdın değerlendirmelerinin ön ve son testte farklı olduğu göz önüne alındığında çok fazla sayıda öğrencinin sınav notunun farklılık gösterdiği görülmektedir. Üstelik iki değerlendirmede de aynı rubriği kullanmış olmalarına rağmen, hocaların çoğunun ilk ve ikinci değerlendirmeleri arasında ciddi farklılıkların bulunması, sadece rubrik kullanımının puanlayıcıların tutarlı olmalarını garanti etmediklerini, dolayısıyla, puanlayıcıların rubriğin doğru değerlendirilmesine dair eğitim almaları gerektiğini savunan birçok çalışmayı desteklemektedir (Chapman & Inman, 2009; Dawson 2009; Gonzales, Trejo & Roux 2017; Kohn, 2006; Reddy & Andrade, 2010 ve Rezai & Lovorn, 2010).

Ayrıca, standardizasyon toplantısına katılmadıklarında, öğretim görevlilerinin öğrenci kağıtlarını çok daha katı bir şekilde değerlendirdiği gözlemlenmiştir. 24 hocadan 17'sinin ikinci değerlendirmedeki notları, birincisine nazaran çok daha düşüktür. Bununla birlikte, her ne kadar Wilcoxon Signed Test veri çözümlemesi sonuçlarına göre, sadece 10 öğretim görevlisinin ilk ve ikinci değerlendirmelerinin ortalamalarının arasında anlamlı bir fark görülmüşse de, tüm kağıtların (n=240) toplam puanları Wilcoxon Signed Test veri çözümlemesi ile incelendiğinde,

kağıtların ilk ve ikinci toplam puanları arasında istatistiksel olarak anlamlı bir fark olduğu görülmüştür. Kısacası standardizasyon toplantılarının puanlama (intra-rater) güvenirliğini arttırdığı bu çalışmanın bulgularından biridir.

Bununla birlikte, standardizasyon toplantılarının puanlayıcı (inter-rater) güvenirliğini de arttırdığı gözlemlenmiştir. Standardizasyon toplantısının düzenlendiği ilk değerlendirmede notları arasında çok farklılık olmamasına rağmen (en fazla 4), ikinci değerlendirmede birçok partnerin notları arasında ciddi tutarsızlıkların olduğu saptanmıştır. Komitenin partnerlerin değerlendirmeleri arasındaki not farkının 5 ve üzeri olduğu kağıtları üçüncü kez okuduğu göz önüne alındığında, standardizasyon toplantılarının zamandan tasarruf ettiren bir yöntem olduğu sonucuna ulaşılabilmektedir.

Ayrıca literatürde, değerlendirmeyi etkileyen faktörlerden birinin puanlayıcıların anadil farklılıkları olduğu tartışılmaktadır. 20 Türk, 4 yabancı öğretim görevlisinin katılımcı olduğu bu çalışmada, anadil farklılıklarının puanlayıcıların değerlendirmelerini etkilemediği sonucuna ulaşılmıştır (Brown, 1995; Hill, 1996; Kim, 2009 ve Zhang & Elder, 2011). Benzer şekilde, literatürde puanlayıcıların yargılarını etkilediği tartışılan bir diğer faktör olan puanlama deneyimi de bu çalışmada incelenmiş, birçok çalışmayla benzer olarak puanlama deneyiminin puanlayıcıların değerlendirmesinde etkili bir faktör olmadığı yargısına varılmıştır (Shohamy, Gordon & Kraemer, 1992 ve Weigle 1999). Ancak, çalışmaya katılan yabancı öğretim görevlisi ve deneyimsiz öğretim görevlisi sayısının az olması nedeniyle, bu bulguları genellemek doğru olmayabilir.

Özetlemek gerekirse bu çalışma, literatürdeki birçok çalışmayla benzerlik göstererek, standardizasyon toplantılarının hem puanlayıcı hem de puanlama

güvenirliğini arttırdığı (Barkhuizen, Knoch, & Randow, 2007; Brown, 1995; Elder, Lunz, Wright & Linacre, 1990; Greatorex & Shannon 2003; Knoch, 2009; Lumley & McNamara, 1995; McNamara, 1996; Robinson, 2000; Shohamy, Gordon & Kraemer,1992; Sweedler-Brown, 1985; Wang, 2010 ve Weigle, 1998) ve puanlayıcıların öznel yargısını gerektiren tüm nitel ölçütlerin değerlendirmesinde uygulanması gerektiği sonucuna ulaşmıştır.

Öte yandan bu çalışmayla ilgili bazı kısıtlamalar söz konusudur. İlk olarak, ikinci aşama sekiz ay sonra yapılmış olsa da, katılımcı öğretim görevlilerinin ilk aşamada değerlendirdikleri kağıtları hatırlama ve verdikleri ilk notlardan etkilenme ihtimalleri olabilir. Ayrıca, çalışmanın ikinci aşamasında, birbirlerinden farklı ve yoğun programları olması sebebiyle, ilk aşamadan farklı olarak, öğretim görevlilerinden kağıtları bir sınıfta toplanarak okumaları istenememiş; bunu yerine kendilerine kağıtları değerlendirmeleri için bir haftalık süre verilmiştir. Dolayısıyla, katılımcıların kağıtları notlandırdıkları koşulların farklı olması, değerlendirmelerini etkileyen faktörlerden biri olabilir. Yine hocaların yoğun programları ve çalışmaya katılımının gönüllülük esasına dayalı olması sebebiyle, 60'tan fazla öğretim görevlisinin görev aldığı Dil Okulunda, bu çalışma sadece 24 hocayla yürütülebilmiştir. Bununla birlikte, katılımcı hocaların her nitel sınav öncesi standardizasyon toplantılarına katılması zorunlu olması nedeniyle, ikinci aşamada bir standardizasyon toplantısına katılmasalar da, hocaların nitel değerlendirme sırasında kendilerinden beklenenlerin az çok farkında olmaları, yargılarını ve değerlendirmelerini etkilemiş olabilir.

Ayrıca, bu çalışmada ölçme ve değerlendirme alanına ilişkin bazı önerilere yer verilmektedir. İlk olarak, değerlendirme sırasında rubrik kullanımı ve ikili

puanlama prosedürüne ek olarak, her yazma becerisi değerlendirmesi öncesinde standardizasyon toplantısı düzenlenmesi gerekmektedir. Ancak, bununla birlikte, standardizasyon toplantılarının güvenirlik üzerine etkisinin yanında verimliliği de göz önüne alınmalıdır (Pufpaff, Clarke & Jones, 2015). İş yükü zaten çok fazla olan öğretim görevlileri, bu zorunlu toplantıları bir külfet olarak görebilir ve gerekli verimi alamayabilirler. Dolayısıyla kurumlar, hocaların iş yükünü hafifletebilmek adına, her zaman yüz yüze buluşmak yerine, online toplantılar ya da forumlar şekliyle de yürütebilecekleri standardizasyon toplantıları düzenleyebilirler.

Ayrıca bu çalışma, çalışmada elde edilen bulgulara ve kısıtlamalara dayanarak, gelecekte benzer konuda yapılabilecek çalışmalara ışık tutacak bazı önerilerde bulunulmuştur. Bu çalışma öğretim görevlilerinin öğrencilere verdiği toplam notlar üzerinden yürütülmüştür. Daha sonra yapılacak çalışmalar, puanlayıcıların değerlendirmelerini detaylı bir şekilde anlayabilmek adına, rubrik üzerindeki her bir barem için verilen not üzerinden yapılabilir. Ayrıca, ileride bu çalışmada kullanılan nicel veri çözümlemesi ile birlikte, hocaların ilk ve ikinci değerlendirmelerini ve neden bu notlarda karar kıldıklarını daha iyi anlayabilmek adına, değerlendirmeler sırasında Sesli Düşünme Protokolü, ve değerlendirme sonrasında bir röportaj uygulanabilir. Bununla birlikte, metodolojide bir değişikliğe gidilerek, daha önce bu tür toplantılara hiç katılmamış bir deney grubu ile, sürekli bu toplantılara katılan bir kontrol grubu oluşturulup, bu tür toplantılar hakkında hiç bilgi sahibi olmamanın puanlayıcıların kararları üzerine etkisi incelenebilir.

E. TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences ☐

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ■

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics ☐

Enformatik Enstitüsü / Graduate School of Informatics ☐

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences ☐

YAZARIN/ AUTHOR

Soyadı / Surname : Karadenizli-Çilingir
Adı / Name : Mahmure Nur
Bölümü / Department : İngiliz Dili Öğretimi

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English): The Effect of Standardisation Sessions Conducted Before English Language Writing Exams on Inter-Rater and Intra-Rater Reliability

TEZİN TÜRÜ / DEGREE: **Yüksek Lisans** / Master ■ **Doktora** / PhD ☐

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ■

2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two year. *** ☐

3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. * ☐

***** *Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir.*
A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.

Yazarın imzası / Signature ..... **Tarih** / Date .....