HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Department of Foreign Language Education
English Language Teaching Program

NATIVE AND NON-NATIVE RATER APPROACHES TO WRITING EVALUATION IN PREPARATORY CLASSES

Ece GÜRBÜZ

Master's Thesis

Ankara, 2024

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

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

HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Department of Foreign Language Education
English Language Teaching Program

NATIVE AND NON-NATIVE RATER APPROACHES TO WRITING EVALUATION IN PREPARATORY CLASSES

HAZIRLIK ÖĞRENCİLERİNİN YAZMA BECERİLERİNİN ANA DİLİ İNGİLİZCE OLAN VE OLMAYAN OKUYUCULARIN YAKLAŞIMLARINA GÖRE DEĞERLENDİRİLMESİ

Ece GÜRBÜZ

Master's Thesis

Ankara, 2024

Abstract

Performance assessment in second language learning has always been problematic in ESL/EFL communities due to its subjective nature. Writing assessment, in particular, is a crucial element in language learning and its assessment. While various methods to enhance reliability in writing assessment have been studied, the approaches of native and non-native raters have not been extensively researched. This study investigates the differences between native and non-native speaker raters' approaches to writing assessment in a state university preparatory class in Turkey. The raters were observed for consistency, severity, and decision-making behaviors, with the first two showing no significant differences. However, in terms of decision-making behavior, NNS raters were found to focus more on organization and use of language, while NS raters emphasized content. The results indicate that native and non-native raters generally do not differ considerably in their approach to writing assessment. They consider similar elements and follow similar steps, with the main differences being NNS raters' critical perspectives, the advantages and disadvantages of being an NS/NNS rater, and the high standards of ELT education in Turkey.

Keywords: performance assessment, second language learning, writing assessment, Native speakers, non-native speakers, rater consistency, rater severity, decision-making behavior

Öz

İkinci dil öğreniminde performans değerlendirmesi, ESL/EFL topluluklarında her zaman bir tartışma konusu olmuştur çünkü öznel bir doğaya sahiptir. Özellikle yazma değerlendirmesi, dil öğreniminde ve değerlendirmesinde önemli bir unsurdur. Yazma değerlendirmesinde güvenilirliği artırmak için çeşitli yöntemler araştırılmış olsa da, anadili İngilizce olan ve olmayan değerlendiricilerin yaklaşımları geniş çapta incelenmemiştir. Bu çalışma, Türkiye'deki bir devlet üniversitesinin hazırlık sınıfında anadili İngilizce olan ve olmayan değerlendiricilerin yazma değerlendirmesine yönelik yaklaşımları arasındaki farkları araştırmaktadır. Değerlendiriciler tutarlılık, katılık ve karar verme davranışları açısından gözlemlenmiş olup, ilk iki konuda önemli bir fark bulunmamıştır. Ancak, karar verme davranışları açısından, anadili İngilizce olmayan değerlendiricilerin daha çok organizasyon ve dilin doğru kullanımı ile ilgilendiği, anadili İngilizce olanların ise içeriğe odaklandığı görülmüştür. Sonuçlar, anadili İngilizce olan ve olmayan değerlendiricilerin yazma değerlendirmesine yaklaşımlarında genel olarak önemli bir fark olmadığını göstermektedir. Her iki grup da benzer unsurları dikkate almakta ve benzer adımları izlemektedirler. Başlıca farklar ise anadili İngilizce olmayan değerlendiricilerin eleştirel bakış açıları, anadilin değerlendirme konusunda sağladığı avantajlar veya dezavantajlar konularında görülmüştür.

Anahtar sözcükler: performans değerlendirmesi, ikinci dil öğrenimi, yazma değerlendirmesi, anadili İngilizce olanlar, anadili İngilizce olmayanlar, değerlendirici tutarlılığı, değerlendirici katılığı, karar verme davranışı

Acknowledgements

First and foremost, my warmest thanks go to my supervisor Assist. Prof. Dr. İsmail Fırat ALTAY who trusted and supported me all through this process. His encouragement and motivating behaviors kept me on track when I felt exhausted. I learned a lot from his advice, feedback and constructive critics on my work. I also owe a great deal to his positive point of view that helped be keep my belief in myself.

I am also grateful to the committee members Assoc. Prof. Ceyhun KARABIYIK and Assist. Prof. Dr. Hatice ERGÜL for their esteemed interpretations and suggestions.

I am equally thankful to TÜBİTAK for their award of the funding and financial support during my studies.

Last but certainly not least, I wish to express my deepest gratitude to all my beloved ones who have supported me throughout this journey. To Aylin, Ünal, and Demir Gürbüz, for unwavering support and patience. To Mithat Yeter, for constant encouragement and belief in me, which have been invaluable. I also extend my heartfelt appreciation to my friends and participants for their contributions and support. The collective encouragement, patience, and assistance have been instrumental in helping me reach this milestone. Thank you all from the bottom of my heart.

Thank you all for being an integral part of this journey. Your love and support have made all the difference.

Table of Contents

List of Tables

List of Figures

Symbols and Abbreviations

ELT: English Language Teaching

ESL: English as a Second Language

EFL: English as a Foreign Language

L2: Second/Foreign language

NS: Native speaker (of English)

NNS: Non-native speaker (of English)

Chapter 1

Introduction

Writing assessment has been a complicated issue in language learning mainly because writing evaluation and scoring depend on personal judgement of individuals that is constituted by their own unique interpretation and background. This leads the subjective assessment and reliability concerns which is also an important subject for results of these tests are regarded as certifications of the language proficiency and these results may play an vital role for the test taker's business or education life. The first chapter of this study includes the statement of the problem giving brief information about the core of the problem and this part is followed by an explanation of the purpose of the study finalizing with the research questions.

Writing assessment has been a complicated issue in language learning mainly because writing evaluation and scoring depend on the personal judgment of individuals who are constituted by their unique interpretations and backgrounds. The subjective assessment and reliability concerns occur because of this. This is an important subject for the results of these tests are regarded as certifications of language proficiency, and these results may play a vital role in the test taker's business or education life. The first chapter of this study includes the statement of the problem. This part provides brief information about the core of the problem and is followed by an explanation of the purpose of the study with the research questions.

Statement of the Problem

Performance assessment in language learning contains the scoring of productive skills such as speaking, and writing. A prominent method to assess performance in productive skills is through direct evaluation of the spoken or written text instead of multiple choice based tests, which are more reliable in scoring since there is no room for raters' interpretations (Hughes, 2003). Writing evaluation and scoring, on the other hand, pose challenges in English language teaching due to its subjective nature, requiring the raters to assess the students' texts personally. This subjectivity can lead to variations in grading, influenced by raters' backgrounds, education, and experience (Horowitz, 1991; Barrett, 2001). Additionally, factors such as the raters' native language can also affect the rating process. Subjectivity in writing assessment is a concept that undermines reliability and

validity because it hinders the ability to provide fair and accurate grades. Even though there is always a room for subjectivity to some extend in direct writing assessment, the variance in practices and approaches can still be reduced and because they should be as reliable and standard as possible for the results may have important role for the attendees' opportunities. Alternatively, they may not provide an accurate reflection of their writing competency.

Aim and Significance of the Study

Native speakers are natural acquirers of a language and Native English speakers are actively working in various language-teaching settings including language departments, preparatory classes and private schools in Turkey, as well as in many other countries. They are also involved in writing evaluation and assessing procedure indeed. Naturally, the perceptions of English language held by a Native speaker (NS) and non-native speaker (NNS) differ. This raises the question of whether their approaches to writing assessment differ significantly and if there are notable disparities in reliability and fairness between NS and NNS ratings in terms of their decision-making behavior and scoring behavior. Decision-making behavior refers to the strategies and principles that the raters follow gradually. This procedure enables to reveal their approach to assessment criteria and their understanding of a high or low quality writing. In addition to decision-making behavior, scoring behavior in terms of severity and leniency in assessment can be investigated and the results can refer to what is important for a good text in their opinion. Spotting these differences and examining the main reasons will help unveil the deficiencies or strong sides of these two groups, reduce the range of alterations in scoring, learn from each other and increase reliability and standardization in writing assessment as it is intended in this particular study.

Research Questions

1. To what extend do Native and non-native raters diverge in their assessment of identical paragraphs?
2. How do Native and non-native raters consider various criteria when evaluating writings?

Assumptions

It is assumed that this study will contribute to the subjectivity and reliability issues in language learning and writing assessment. Specifically, it seeks to uncover the underlying

reasons of the alterations in scoring differences of native and non-native perspectives. It will hint at the conflicting question of whether the NSs' judgement is more reliable and accurate than that of the non-natives.

The study will depend on both qualitative and quantitative data. Qualitative data will be the majority which primarily consists of the rater's think-aloud sessions. Through this method, it is expected to reach a detailed stream of ideas and interpretations reflecting the main concerns. As for the quantitative part, the rater's scores are to be analyzed through descriptive statistics to identify significant similarities and differences and any meaningful fluctuations. The quantitative data is assumed to support the qualitative.

Raters will evaluate two groups of writing samples both holistic and analytic methods. The writing samples sourced from a state university, will be equally distributed for evaluation. They will be assessed by a NS and NNS group of raters. The first group of texts will undergo impression scoring, relying on raters' innate judgment without a predefined scale. This method will be helpful to recognize their pure and absolute considerations allowing for a wide range of ratings. The second group of texts will be rated using a rubric. This part would be functional to look into how they employ the criteria, and how they perceive them.

It is anticipated that the results give clues about the variance in the rater's severity, consistency in decision-making behavior, and the efficacy of the scoring method. Additionally, part, the participants will be asked their ideas about the abovementioned issues to gain an in-depth grasp of their consideration.

Limitations

This research has limitations as every other research does. To start the participants, were in two groups NS and NNS raters. 20 English Instructors participated 10 for each group. All Natives were people who currently work or once worked in several private and state universities' preparatory classes in Turkey. However, the participant demographics lacked sufficient diversity, so their backgrounds were not evenly distributed such as education level, gender, or educational experience. All the NNS raters, except for only one, are employed in the same institution where this study was conducted. For this reason, NNS participants may have shared the institutional assessment tradition and ways of rubric use. Additionally, the background education and years of experience cannot be distributed equally. NS raters were more experienced than the non-natives in terms of year of teaching were. Thus, factors such as the experience- and background of the raters were not the

scope of the study. However, there are studies in the literature that has reviewed these factors (e.g., Santos, 1988; Vann, Lorenz, & Meyer, 1991). With a larger group, a more diverse set of parameters could have been correlated. The number of participants was too small to generalize the results. Furthermore, it was not possible to train the subjects before and after the research and it was based on a single exam sample, something more longitudinal could be designed (e.g. to see if they would change their behavior after receiving training or working together). The native group assesses the texts via Zoom meetings on the computer but the native group could have a chance to take notes on sheets. The meetings lasted more than one hour. This length and the time limitation caused a narrow number of the participants were be able to involve. Finally, this study was done without the groups seeing each other. A study can also be conducted before and after they share ideas.

Definitions

Test: A systematic procedure for observing a person's behavior and describing it with the aid of a numerical scale (Cronbach, 1971).

Testing: the act of determining and assessing student learning with particular practices (Harris & McCann, 1994).

Assessment: the measurement of performance and its representation through a numerical value (Brown, 1996).

Performance assessment: It depends on the test-taker's level of expertise, knowledge, and general performance productive skills, written or spoken output. (Norris et al., 1998).

Evaluation: Collection and perusal of information to make decisions about people (Bachman, 1990).

Reliability: The test's score is consistent when it is administered multiple times (Brown, 2004).

Validity: The idea that a test should aim to test the things it intends to test (Harmer, 2004).

Scoring / Rating: The process of grading process (Weir, 2005)

Holistic Scoring: The scoring method regarding grading the students' performance according to a set of predetermined criteria that evaluates the texts globally (Brown, 1996)

Impression scoring: The holistic scoring method that requires the raters to award one overall grade to the text according to their criteria, without any rubric or predetermined criteria (Hughes, 2003).

Analytical Scoring: The method of scoring regarding the rater awards a grade to the text through a set of predetermined criteria presented on a rubric (Brown, 1996).

Rubric: A table that shows the predetermined descriptors of the assessment criteria (Harris & McCann, 1994).

Native (English) Speaker: Someone who acquires a language and has spoken it since infancy as their mother tongue (Cambridge Dictionary n.d.). Particular to this study, Native English Speaker (NS) refers to the natural speakers of English.

Non-Native (English) Speaker: Someone who has learned a particular language as a child or adult rather than as a baby (Cambridge Dictionary n.d.). Particular to this study, non-native English speakers refers to those who learned English as a foreign language (EFL) or English as a second language (ESL).

Conclusion

This study analyses the challenges in writing assessment focusing on the NS and NNS raters. By observing their approaches, it aims to understand the methods and strategies they use and reveal the differences between these two group. The findings are expected to provide data about the details in writing assessment. The implications are to be of importance since writing assessment has been a challenging area even though some limitations occurred during the time of research.

Chapter 2

Literature Review

Introduction

Assessing writing skills in language learning is complex due to the intricate cognitive processes involved in creating well-organized texts. Mastery of linguistic, structural, and cultural elements is necessary, alongside the influence of the writer's background, resulting in a unique, personal output. Writing assessment considers both the individual's work and the rater's background, including experience, native language, and education, leading to subjective results. Despite the inevitability of subjectivity, it is crucial for ensuring reliability, especially for assessments influencing employment or educational opportunities. Scholars have recognized the issue, attributing subjectivity and reliability problems to scoring variance, rater consistency, and behavior.

The Nature of Writing Skill

In today's information age, effective writing has become an indispensable means of communication globally. Interactions among nations have intensified due to cultural, social, educational, and business reasons, facilitated by advancements in technology and transportation. This surge in interactions has not only allowed nations to engage but has also mandated the development of efficient ways to convey crucial information. As a result, as one of the four primary skills of language acquisition, writing skill has been studied to find the best possible way to develop and assess. To be able to understand the nature of writing skill, we can compare and contrast it with the other skills.

In language acquisition, there are four primary skills: reading, listening, speaking, and writing. Two of these skills are considered receptive, involving the examination and comprehension of meaning from written or spoken texts (Harmer, 2007). Conversely, speaking and writing are categorized as productive skills, requiring the language user to actively generate meaning through verbal or written expression based on their linguistic knowledge. In essence, in receptive skills, the language user functions as a decoder, while in productive skills, they operate as an encoder (Brown, 2006). The spotlight on productive skills, particularly ESL/EFL writing and speaking (Weigle, 2002), underscores the significance of writing as a pivotal aspect of language learning and assessment.

When compared to speaking in terms of their practice and process, these two skills sometimes differentiates vaguely from each other. However, there are times that the distinction between them is very much obvious (Harmer, 2004). For instance, in text

messaging or chatting, writing is more akin to speaking because the text is produced instantly as thoughts arise, without any attempt at editing. On the other hand, spoken lectures are more comparable to written texts as the content is derived from scientific texts, following a formal and predetermined pattern. However, the distinction becomes evident when considering accuracy and permanence. Weigle (2002) argues that language accuracy is more crucial in written texts than in spoken ones, and they are highly esteemed in education. Harmer (2004) emphasizes the permanence of writing and the transience of speaking, stating, 'Spoken words fly away on the wind; written words stay around, sometimes, as we have seen, for hundreds or thousands of years' (p. 7). As one of the four essential language skills, writing has evolved into a key activity for cross-language communication. Consequently, the study of writing teaching and assessment has emerged as a vital topic in language sciences to meet the demands of this evolving linguistic landscape. Grabowski (1996) summarizes its importance as follows:

> Writing as compared to speaking can be seen as a more standardized system, which must be acquired through special instruction. Mastery of this standard system is an important prerequisite of cultural and education participation and the maintenance of one's rights and duties. (Grabowski, 1996: 75).

The definition of writing skill has also changed and multiplied since the beginning of its study. The earliest way that has been recorded in literature is merely recording speech implementing particular grammatical or lexical structures. However, it was later when it is understood that a well-developed written text is hard to produce not only in one's first language, but also in their second language (Brown, 2003). Referring to this complexity, it is described as a "non-linear" activity in which the writers attempt to regenerate an approximate information in their mind using several "micro and macro skills" (Zamel, 1983, Casanave, 2004). According to Heaton's (1990) outline, these skills include several key elements that constitutes main criteria of rubrics: precision in language use and mechanics, encompassing correct spelling and punctuation; the construction of content through powerful and relevant ideas; and the organization of the essay with a coherent and fluent narration (p. 135). As a result of these aspects, it can be clearly stated that it is a cognitive activity and the writer's mental performance can be resembled to a "problem-solving activity (Hyland 2010).

As an increasing number of studies are conducted, it has become evident that the nature of writing skill is intricate and sophisticated, making it impractical to provide a comprehensive umbrella term to define it. Instead, it is more feasible to delineate its significant aspects individually (Weigle, 2004). Upon closer examination of its definition, it

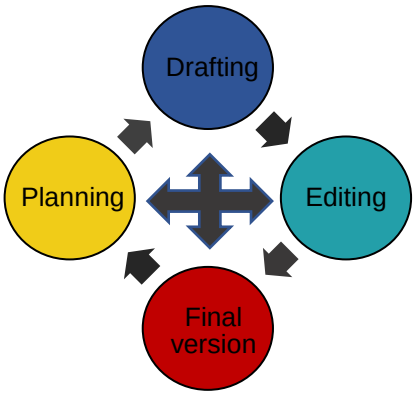
becomes apparent that writing skills are linked to academic achievement and more standardized compared to speaking (Grabowski, 1996). Moreover, they are regarded as an integral part of the curriculum and a tool for learning and expanding one's knowledge by articulating existing understanding (Weigle, 2004). Additionally, writing serves as a reflection of culture, as it necessitates an understanding of the way of life of people within that culture (Harmer, 2004).

Writing process has particular stages according to scholars. It is an intricate process that encompasses thinking, planning, organizing, and linking, involving various levels of language manipulation—both at the sentence and clause levels, as well as at the word and phrase levels—along with considerations for spelling and punctuation (Hamp-Lyons, 2003, Harmer, 2004).

Harmer (2004) proposes a clear rationale for the staging of the writing process. The planning stage is dedicated to establishing the writing's objectives, analyzing the intended audience, and outlining the main ideas to form an overall content structure. The first draft then takes shape by articulating these ideas into statements, which are subsequently reviewed and edited. The writer revises the statements with a critical perspective, addressing any issues related to ambiguity, linguistic and lexical accuracy, and organization. Prior to the final version, the writer may engage in multiple rounds of re-planning, re-drafting, and re-editing. However, due to the complexity of the writing process, the writer may move among these stages in a non-linear manner, akin to a "stream of consciousness" (Harmer, 2004)

Figure 1

The model depicting the stages of writing process (Harmer, 2004, p.6)



One other important component that is necessary to be considered is genre of the written text. Genre refers to the classification of texts according to the current topic and common ways of how to respond to that particular situation and each genre has its own

linguistic and structural characteristics. These characteristics also effects the purposes of writing and its assessment accordingly (Hyland, 2010). Some are shown in the following figure.

Figure 2

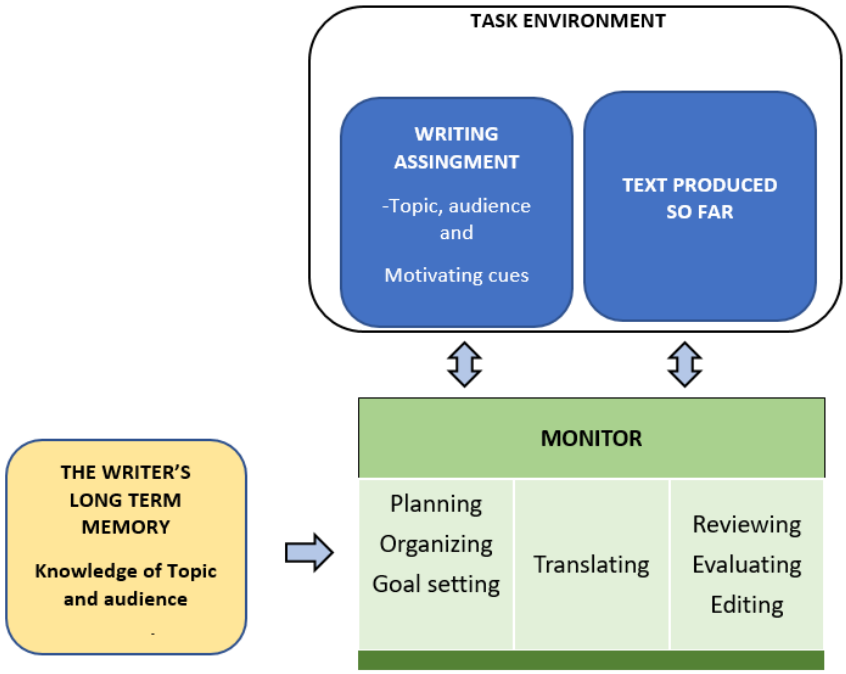
Main types of genres (Brown, 2004).

1. Academic Writing
Papers and general subject reports Essays, compositions Academically focused journals Short-answer test responses (e.g. lab reports) Technical reports Theses, dissertations
2. Job-related writing
Messages (e.g. phone messages Letters, emails Memos Reports (e.g. job evaluations, project reports) Schedules, labels signs Advertisements, announcements
3. Personal writing
Letters, emails, greeting cards, invitations Messages, notes Calendar entries, shopping lists Financial documents (e.g. checks, tax forms, loan applications) Diaries, personal journals Fiction (e.g. stories, poetry)

According to Hayes & Flower, writing skill functions cognitively in three main processes in terms of linguistic performance. The writer first establishes an objective of writing and organize the main ideas, then exploits his/her long term memory to call prior knowledge about the topic and revise the product respectively and this is how we write simply (1980). In other words, writing is a lively, intellectual and cyclical process when the writer thinks critically and highly creatively (Hyland, 2010). As Weigle states, writing proficiency points out student's potential of success and developed analytical thinking ability (2002). As a result, among the four skills, writing is considered one of the toughest to excel (Hamp-Lyons 2019).

Figure 3

The Hayes & Flower (1980) writing model



The reason why it is a difficult skill to master is that writers need to be competent in several domains. That is to say, they need to activate not only their linguistic knowledge, which is about accurate language structures, but also their knowledge of discourse and sociolinguistics (Hymes, 1972; Canale and Swain, 1980; Bachman, 1990; Grabe & Kaplan, 1996). To summarize, Grabe and Kaplan depicted a taxonomy of the necessary language knowledge (1996). For them, linguistic knowledge includes formatting a piece of writing, spelling, and punctuation rules. Additionally, linguistic knowledge encompasses choice of vocabulary and syntax. Sociolinguistic knowledge helps the writer to choose the appropriate and acceptable ways of using a language in a particular social atmosphere, and discourse knowledge helps them with the unity of a written piece. Briefly, suitable use of functional language, the audiences' situation, and formality degree are involved in this part. On the other hand, discourse knowledge deals with genre structure and constraints, use of conjunctions to clarify the meaning, and overall organization (Grabe & Kaplan, 1996). In other words, besides knowing the use of language, a writer should be knowledgeable about the harmony of the language and the communicative traditions of people (Weigle, 2002).

In addition to its cognitive nature, writing serves as a social and cultural phenomenon. Specific features contributing to the favorability of a text can vary across communities (Hyland, 2010). For example, English prose tends to be more 'transparent' and

'explicit' compared to Eastern prose, where the main idea is often implied. Additionally, English prose is characterized by hierarchical organization and directness, contrasting with the Spanish style that allows for longer introductions rather than concise statements, or the Chinese style, which favors numerous examples over direct statements. Considering these differentiations, it can be concluded that writing a well-organized piece of text or essay is actually related to meeting the readers' expectations that can also change depending on their culture or personal ideas. (Ostler, 1987; Yorkey, 1977; Kaplan, 1966; Leki, 1992 as cited in Weigle, 2002)." Thus, the writers also need to discover what is acceptable for the readers rather than planning or writing the text. In support of the importance of evaluating paragraphs, Hayes (1996) emphasized that reading with the intent to assess a written work holds crucial significance. It allows for the detection of deficiencies and facilitates the refinement of the piece according to the readers' preferences if necessary. With this purpose in mind, Hayes presents the cognitive process of a writer, as depicted in the following figure.

Figure 4

Cognitive process in evaluative reading (Hayes, 1996).

Possible Discovery	Comprehend & Criticize	Possible Problem Detection
new diction	decode words	spelling faults
alternative construction	apply grammar knowledge	grammar faults
puns and alternative interpretations	apply semantic knowledge	ambiguities and reference problems
new evidence and examples	make instantiations and factual inferences	faulty logic and inconsistencies
analogies and elaborations	use schemas and world knowledge	errors and schema violations
ideas for alternative text structures	apply genre conventions	faulty text structure
ideas for transitions and connectives	İdentify gist	incoherence
alternative plans	infer writer's intentions & perceptions	disorganization
new voice and alternative content	consider audience needs	inappropriate tone or complexity

Briefly, in the act of writing, authors engage with three essential language domains: linguistic, sociolinguistic, and discourse knowledge. Linguistic knowledge concerns language structure, encompassing syntax and grammar. Sociolinguistic awareness involves navigating social conventions in language use. Discourse knowledge focuses on stylistics and narrative harmony. Effectively employing information from these domains is

crucial for writers aiming to craft socially appropriate compositions. Being merged into other language skills, writing can be considered both as a tool and as a sign for language development, which also brings the function of assessment into prominence.

Key Concepts in Assessment

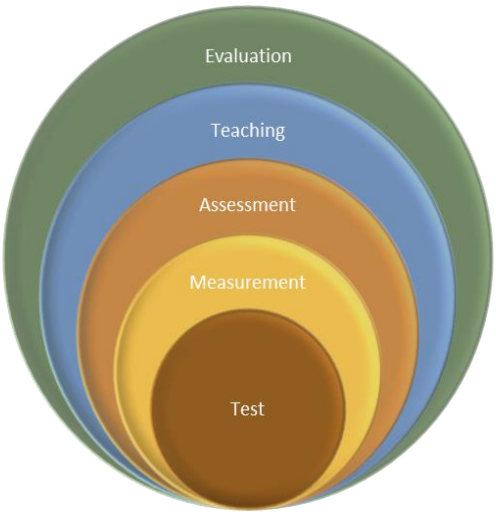
Before delving into the specifics of the study, it would be beneficial to elucidate fundamental concepts related to the subject matter for the reader's enhanced comprehension. Defining overarching terms such as "testing," "assessment," "evaluation," and "measurement" would serve as a suitable starting point for they are sometimes perceived as synonyms though they are basically different concepts. First, in the context of language acquisition and other domains, individuals undergo "testing" to demonstrate their proficiency in a given area or their capacity to perform a specific task. Essentially, tests are regarded as instruments utilized to ascertain an individual's capability or to make informed judgments regarding their performance in a particular domain. It is imperative that test results are not arbitrary, but rather adhere to a fair and widely accepted framework (McNamara, 2000; Fulcher, 2010). More specifically, Brown defines it as “a method of measuring a person’s ability, knowledge or performance in a given domain” (2004). Therefore, the process of testing is inherently linked to quantitative measures or numbers, as it involves “measuring” individual or entity to determine their proficiency, expertise, or achievements to place them or diagnose their learning improvement (Casanave, 2004, Brown & Abeywickrama, 2010, Brown, 1996).

Assessment and evaluation are distinct concepts, although they are often used interchangeably. The term "assessment" encompasses a broader scope, encompassing not only formal and structured tests, but also informal methods such as commentary and spontaneous feedback from teachers to students (Brown & Abeywickrama, 2010). In contrast, evaluation is a more comprehensive term, involving interpretation and judgment rather than mere rating (Baxter, 1997). While assessment refers to the formal and informal processes of measurement typically conducted by teachers to gather information and guide instruction, evaluation involves deeper interpretation and consideration of learning and materials (Harris & McCann, 1994). Bachman further distinguishes these terms, defining measurement as the process of quantifying and tests as the instruments used for this purpose, while evaluation entails making judgments about performance and defining the abilities of the test subject (1990). Fachrurrazy provides a concrete example to illustrate these concepts, highlighting that the scores themselves hold no meaning until they are

interpreted and used to make judgments about the students' performance. This relationship between assessment and evaluation is depicted in the following figure.

Figure 5

Relationship about the terms (adapted from Brown & Abeywickrama, 2010)



Writing Assessment and Subjectivity

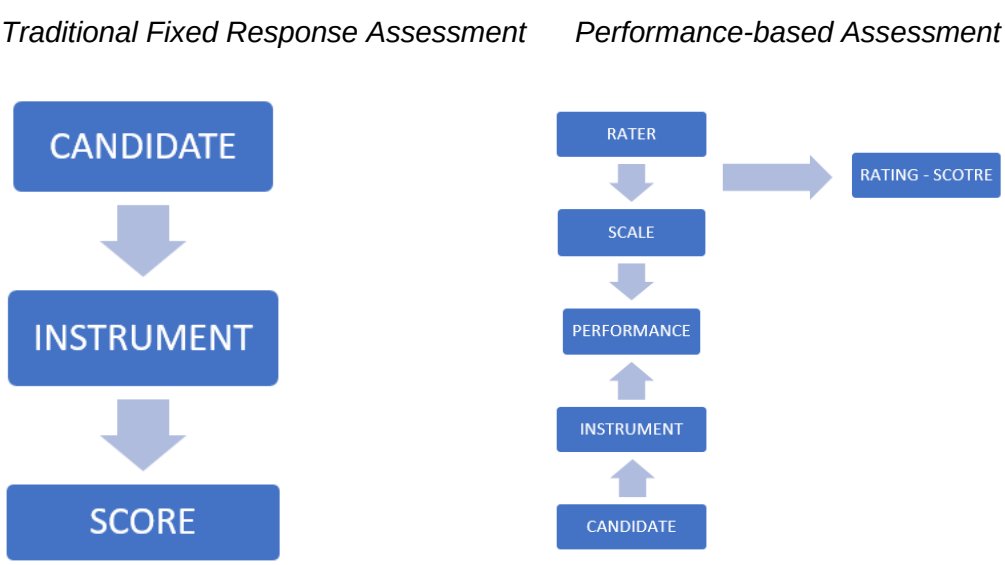
As productive skills, writing and speaking are performance-based skills and assessment of productive skills can be also labeled as performance assessment because it depends on test taker proficiency, knowledge and overall performance in spoken interviews and essay writing (Norris et al., 1998). In language learning, performance assessment depends on learner's spoken or written production. It is distinguished from traditional tests since it includes a performance of the test-taker in a particular subject and the quality of the performance is determined through an "agreed judging process" (McNamara, 1996). Even though it is a demanding process for the raters, the test-takers can present their genuine capability in an area because it deals with real-life situations and problems and the tasks can consist of portfolios, open-ended questions or essays. In comparison to traditional tests, they are more functional to demonstrate the test takers' proficiency (Fitzpatrick & Morrison, 1971; McNamara, 1996; Brown & Hudson, 1998).

Earlier writing tests are designed as standardized multiple-choice items towards the learners' language use or mechanics competence. The underlying rationale here is to conduct 'implicit testing', also called an 'indirect testing'; which means testing the sub skills of writing rather than having the test takers perform their proficiency in grammar and vocabulary use, mechanics, capitalization, spelling etc. by writing a paragraph. Indirect

method was favored for their perceived objectivity. Over time, this strategy has evolved and a shift occurred with the introduction of direct assessment methods, which focus on evaluating individual performance. This transition reflects a deeper understanding of the nuances involved in assessing writing skills (Baxter, 1997; Hatipoğlu, 2022; Huot et al., 2010; Brown, 2004; Casanave, 2004, Brown & Hudson, 1998). Today, it is widely believed that the test-takers should have the opportunity to demonstrate their original proficiency by performing even though the process is long and the product is difficult to rate because "getting the people to write is the best way to test their writing ability" (Brown & Hudson, 1998; Hughes, 2003).

Figure 6

The Characteristics of Performance Assessment (Kenyon, 1992; McNamara, 1996)



Mastering in performance assessment has its own difficulties since subjectivity and difference in rater interpretation and scores may hinder reaching reliable results. However, it is also a crucial skill as its "every teacher's job" (Hamp-Lyons, 2003). Specific to writing assessment, it involves evaluating and assessing an individual's writing proficiency by assigning a numerical rating to their work based on subjective judgment (Barrett, 2001). One of the primary problems with writing assessments is the issue of subjectivity, which refers to the fact that different raters will assign different ratings to the same essays. Numerous studies have addressed this issue and offered solutions even though its complete elimination is generally accepted impossible because performance cannot be

assessed precisely with a definite answer key (Attali, 2016; Bachman, 1990; Wolcott & Legg, 1998, Eckes, 2005; Kayapınar, 2014; Kondo-Brown, 2002; Meier, 2012; Moskal, 2000; Schaefer, 2008; Vaughan, 1991). Elbow (2003) points out the dilemma plainly: “To rank reliably means to give a fair number, to find the single quantitative score that readers will agree on. But readers don’t agree” (p.188).

In contrast with performance assessment, multiple-choice tests are relatively objective in comparison with marking a written work, which is unreliable because individuals practice it (James 1977; Attali, 2016). One additional dilemma arises when judging the quality of writing skills, particularly in domains such as grammar and vocabulary. While widely accepted that writing is a complex activity, encompassing elements like meaning and rhetoric beyond linguistic forms, assessments often grapple with the challenge of adequately capturing this complexity (Casanave, 2004).

A common practice in today's assessment tactics is to mitigate this subjectivity by having writings assessed by two different raters. In addition, the raters are trained to grow an objective approach and in some institutions, the variance in raters’ evaluation process is tried to be reduced through standardization meetings. Still, the raters are not tend to agree on their judgements and that reduces interrater reliability. Some scholars have the opinion of accepting the subjective nature of writing assessment, to some extent, as it is, and it is not sensible trying to base it upon a very objective mathematical system (Casanave, 2004; Abedi, 2010; Rezaei & Lovorn, 2010). As Hamp-Lyons (2003) also implied, writing is an elaborate production for both the writers, and the raters cannot be apart from their humane characteristics and background. Thus, the writings and the ratings are inevitably influenced by their education, gender, culture, emotional state, characteristics, nationality, language ability, or character. No matter the strategy developed, writing assessment never possesses the characteristic of complete systematicity or conciseness, as both the producer and assessor of the text are human beings who constantly change (Hamp-Lyons, 2003).

Thus, performance assessment tasks are subject to the critical perception of individuals constituted by several factors such as their age, experience, culture, gender and character. Previous research that has been concerned about objectivity issues about performance evaluation has found that individualism has led reliability and validity issues occur in this domain (O’Sullivan, 2000; Shaw & Weir, 2007; Brown, 1995; Cumming, 1990; Johnson & Lim, 2009; Mehrens, 1992, Henning, 1996, McNamara, 1996, O’Loughlin (2002).

Reliability, Validity and Consistency in Writing Assessment

Validity and reliability are two crucial test-related concepts. We are unable to discuss fairness or reliable outcomes due to their shortcomings. We can go deeper into these concepts to understand their significance.

Reliability refers to the consistency of a test and stability of the grades. In other words, a reliable test results in consistent or stable scores no matter how many times it is conducted on the same group (Harris and McCann, 1994; Wolcott & Legg, 1998). Brown, 1996; Baxter, 1997, Heaton, 1990; Hamp-Lyons, 2003). That is to say, if a test is reliable, it should end up with similar results in the case of being applied to two different groups with similar characteristics (Brown, 2004). To put it more plainly, a test is reliable if it results in similar score even if it is conducted with the same group regardless of different times and places (Hughes, 2003). Spearman (1904) explains that the reliability of an assessment can be determined by calculating the ratio of true score variance to total score variance. A result close to 1.0 indicates high reliability, while a result further from 1.0 indicates high scoring variance and low reliability.

One of the most significant indicators of the reliability of a test is the consistency among the scores. However, if a test involves open-ended items or interpretative sections such as essay writing, some problems may occur in reliability because the rater's personal judgment effects the scoring procedure (Fulcher, 2010). Different raters typically score the same writings differently and these variation arises from rater's characteristics such as their cultural background, native language, experience even their personal disposition. Conflict is that ideal assessment should be fair, accurate and reliable but these factors are impossible to be fulfilled completely in such a subjective area as performance assessment (Horowitz, 1991). Hamp-Lyons indicates that while reliability is a crucial factor and can be achieved at a remarkable rate, achieving a level exceeding 80% is rare, especially in writing assessments, due to rating variance. With the latest strategies or writing assessment that are developed after years of research and study, we can achieve 75% reliability, when at least two raters score a paper (2003).

Even though it is unrealistic to claim or expect to neutralize the scoring variance when it is done by different individuals, it is an undeniable fact that increased variability in the assessment of texts leads to a less fair and reliable evaluation and that situation may threaten reliability and validity factors (Huot, 1990; Bachman, 1990; Gamaroff, 2000). Thus, rater reliability is also an important factor that can be fulfilled to a good extent and that affects the overall reliability of a writing test (Huang, 2009; Cumming, 1990; Engelhard,

1994; Erdosy, 2004; McNamara, 1996; Pula & Huot, 1993; Santos, 1988; Vaughan, 1991; Weigle, 2002).

The two components of rater reliability—intra- and inter-rater reliability—are closely related to the consistency of the scoring system and the rating procedure. The former speaks about the differences in scores among raters who are part of the same group and have comparable attributes. Put another way, we can speak of inter-rater reliability when most or all of the raters assign the same test a similar score. According to Huot, this phrase also emphasizes the value of reliability and makes reference to evaluation fairness in the following ways:

> ” If readers of student writing do not agree on scores at a high enough rate for the writing they read, then a student score would depend on who read the writing rather than who wrote it (Huot et al., 2010: 499).

As to intra-rater reliability, it is more about a raters' consistency of scoring within itself. In that, a rater should score several essays similarly if their quality is also similar. (Connor-Linton, 1995; Brown, 2004; Cumming, 1990; Alderson et al., 1998; Coffman, 1971; Ebel & Frisbie, 1986; Engelhard, 1994; Alderson Erdosy, 2004; Freedman & Calfee, 1983; McNamara, 1996; Pula & Huot, 1993; Santos, 1988; Vaughan, 1991; Weigle, 2002; Barkaoui, 2008). Consistency is calculated by the rater's agreement with his/her pattern of judgements as well as with those of the other rater's. If the agreement surpasses the average, the raters are deemed as “overfitting” whereas a decreased agreement indicates a situation of “misfitting”. (Hill, 1996, Brown, 1995). While assessing performance, if inter-rater and intra-rater consistency can be fulfilled, to some extend if not totally, then the test may come up with reliable results (Brown, 1996).

Validity, on the other hand, is the extent to which the empirical evidence aligns with the rationale for using test scores (Messick, 1989). In simpler terms, a test is considered valid if it accurately measures what it is intended to measure and if the assessment is invalid it causes a feeling of unfairness (Fulcher, 2010; Brown, 2004; Goodman & Carey, 2004; Harris & McCann, 1994; Johnson & Svingby, 2007; Huang, 2009). If a student executes a strong writing performance, they would be identified as a proficient writer and receives high scores, whereas those with the lower scores are thought to be incompetent in writing and the variations in these scores causes a decrease in validity (Huot, 1990). Similar to reliability, validity is also essential for all tests and writing assessments because it ensures that there is no bias in the evaluation process (Huang 2009). Rating variability is influenced by the systematic rating variance arising from the diverse characteristics of raters, which is why the topic of variance is needed to be discussed (Myford & Wolfe, 2003; Schaefer, 2008).

The main types of validity are content, criterion, face and construct validity. When the topic of the task remains in the purpose domain of the test, that test ensures content validity (Huot, 1990; Hughes, 2003; Brown, 2004). Criterion validity is related to the correlation between the measurement tool and the scores awarded according to a predetermined criterion. Performance assessment requires pre-established criterion to be assessed and face validity concerns the convenience of the items to the purpose of the test (McNamara, 1996, Brown 2004). Construct validity, which is a major component of the tests, is about how well a test measures what is claims to measure (Weigle, 2002). Hamp-Lyons employs a metaphor involving shooting arrows to elucidate the concepts of reliability and construct validity, aiming to solidify the reader's understanding:

> "If you shoot three arrows and they all hit the same spot, your aim is very reliable. But suppose those three arrows hit the edge of the target? Your aim is reliable, but not effective. Suppose one arrow hits the bullseye and the other two go in different directions? You have some effectiveness but no reliability. Now suppose all three arrows hit the bullseye? Your aim is both effective and reliable. Construct validity demands that you not only hit the same spot most of the time, but that it's the right spot." (Hamp-Lyons, 2003).

The reliability and validity condition of a test define its effectiveness. While assessing writing, raters need to consult their personal subjective judgement, which creates variance in scoring and spoils reliability and validity indeed. That is why; scoring variance causes "measurement error". As mentioned above, standardization and rater trainings help to fortify rating consistency (Barkaoui, 2010).

Scoring Methods

Subjectivity in writing assessment has led us to find a way to minimize its negative effects, if not eliminate them. Because of this urge, the raters has been directed to evaluate the texts similarly according to certain principles. Thus, rating scales have been introduced to use while rating (Casanave, 2004), defined as "a tool to describe language proficiency" according to "a series of constructed levels" and the learner's language performance is judged through these criteria (Davies, 1999).The main aim is to construct the assessment onto a solid ground. It is accepted that the set of criteria should be specified to be more objective in judgement and this is how rating scales originated for this purpose (Bachman & Palmer, 1996; Weigle, 2002). Common European Framework of Reference (CEFR) is one of the best indicator and guide for the preparation or the writing assessment rubrics. In

the framework, students' needed proficiency for each level is described in detail for different text types CEFR, 2001).

Table 1

Overall Written Production Scale (CEFR, 2001)

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

Table 2

CEFR Reports and Essays-Illustrative sub-scale

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

Basically, two types of scales are adopted that are holistic and analytic scoring. In holistic scoring, raters give one grade to the text as a “whole entity”, whereas they assess the text according to pre-determined sub-domains such as content, grammar, vocabulary, and organization, and a separate score is given to each domain (Hamp-Lyons, 1991; Goulden 1992, Alderson, et al., 1995; Weigle, 2002, Brown 2004, Barkaoui, 2010, Fulcher, 2010). Using a rubric is regarded the most practical way in this case and they are widely used both as holistic and analytic scale (Johnson & Svingby, 2007). Another type of holistic scoring is impression scoring. Specifically, the rater does not employ a specific scale; instead, they evaluate the text based on their own knowledge and opinions, which may have an impact on the text's consistency and dependability because the rating is based on the opinions of the raters. One marker's definition of "outstanding" may differ from another marker's definition of "good," according to Green & Hawkey (2012) (p. 299).

In performance assessment, the evaluation process is to some extent, the rater's implicit judgement. To be objective as possible, the rubrics should be designed elaborately. Instead of general scales such as "good, average, bad or superior", behaviorally anchored rating scale where there are clearer definitions of the criteria can be more useful (Slater, 1980). A sample for each rubric can be seen below:

Table 3

Holistic Rubric Sample (Wolcott & Legg, 1998)

Scores	Holistic Scoring Scale
Extremely Proficient (6)	Papers receiving scores of 6 generally have abundant, good details. The papers show style and thought, and often there is a strong sense of the writer. These papers have few errors, as the writers seem in command of sentence structure and mechanics.
Proficient (5)	The 5 papers are also detailed and developed with some sense of the writer showing through. The writers seem to understand sentence construction although problems with grammar and spelling can begin to arise.
Moderately Proficient (4)	The 4 papers usually have a thesis developed in some significant way with support, although some papers begin to lose focus, and they are not as detailed as the 5's and 6's. Usually there is a sense of sentence construction even though it is not too sophisticated. Sometimes paragraph problems begin to appear.
Slightly Deficient (3)	The 3 papers provide a clear picture of the subject or a sense of the writer, but they are developed with generalities. Grammatical, spelling, and sentence errors begin to dominate the papers.
Deficient (2)	The 2 papers either have very limited and weak development and some grammatical/ mechanical errors, or they attempt some development and are full of errors.
Seriously Deficient (1)	The 1 papers are extremely short with virtually no development at all. (In a few instances, I's may be given for off-topic papers in which students did not understand the topic at all.)

Table 4

Analytic Rubric Sample (adapted from Andrade & Mycek, 2010)

	4	3	2	1
Content	The paper clearly states an opinion and gives 3 clear, detailed reasons in support of it.	An opinion is given. One reason may be unclear or lack detail.	An opinion is given. The reasons given tend to be weak or inaccurate. May get off topic.	The opinion and support for it is buried, confused and/or unclear.
Organization	The paper has a beginning with an interesting lead, a middle, and an ending. It is in an order that makes sense. Topic and closing sentences and main ideas.	The paper has a beginning, middle and end. The order makes sense. Paragraphs are indented; some have topic and closing sentences.	The paper has an attempt at a beginning and/or ending. Some ideas may seem out of order. Some problems with paragraphs.	There is no real beginning or ending. The ideas seem loosely strung together. No paragraph formatting.
Vocabulary	Descriptive words are used ('helpful' instead of 'good' or 'destructive' instead of 'bad').	The words are mostly ordinary, with a few attempts at descriptive words.	The words are ordinary but generally correct.	The same words are used over and over. Some words are used incorrectly.
Sentence & Fluency	The sentences are complete, clear, and begin in different ways.	The sentences are usually correct.	There are many incomplete sentences and run-ons.	The essay is hard to read because of incomplete and run-on sentences.
Mechanics & Grammar	Spelling, punctuation, capitalization, and grammar are correct. Only minor edits are needed.	Spelling, punctuation and caps are usually correct. Some problems with grammar.	There are enough errors to make the writing hard to read and understand.	The writing is almost impossible to read because of errors.

Holistic and analytic scoring methods are developed with the intention of making a reliable and valid judgement of a written piece and each employs the strategy of giving a chance to the rater to exhibit their comments. In other words, we are supposed to know

about the reasons why the score is awarded to be able to speak of reliability and validity. (Connor-Linton, 1995).

These methods have their own advantages and disadvantages in terms of practicality and reliability as mentioned in many studies in literature (Bachman, 1990; Alderson, 1991; Hamp-Lyons, 1991; Song & Caruso, 1996; Weigle, 2002; Gonzalez, 2017; Johnson & Svingby, 2007; Huang, 2009; Barkaoui, 2010; Rezaei & Lovorn, 2010). Some of them puts forward that holistic scoring has a higher level of validity. However, it has problems with reliability and faces challenges such as “bias, fatigue, internal lack of consistency, previous knowledge of the student, and/or shifting standards from one paper to the next” (Perkins, 1983), and it is not likely to explain why those scores are given (Hamp-Lyons, 1995). Because raters have more freedom to make their own decisions, it is also more subjective than the analytical method. Important factors, for example, could differ among raters. As Goulden indicates (1994) “"the rater may include traits not listed and use personal judgment to determine how important a specific trait is to the overall score" (p. 74). Furthermore, their physical or emotional state may have a favorable or bad impact on the process, or they may overlook certain significant mistakes or shortcomings if they are overly pleased by a single aspect (Hamp-Lyons, 1991; Vaughan, 1991; Kayapınar, 2014). Numerous studies have been conducted on the subjectivity and holistic scoring relationship. The results indicate that grammar is the most essential criterion (Lee, 2009), and that the importance of rating criteria depends on whether the rater considers language or content (Sakyi, 2000). Some has demonstrated that there is a considerable variation in scoring (Hamp-Lyons, 1991; Weigle, 2002). On the other hand, it is rater and student friendly because it is much more affordable, time saving and small errors can be ignored easily (Wolcott & Legg, 1998; Johnson & Svingby, 2007; Weigle, 2002).

Analytic scoring, unlike holistic scoring, offers higher inter-rater reliability because it allows us to evaluate student performance in each criterion separately. Studies indicate that using an analytic rubric can increase reliability even more than years of teaching experience (Gonzalez, 2017). An analytic approach breaks down evaluation into specific domains and assesses the text according to predetermined principles for each criterion individually. These domains typically include well-organized content, effective vocabulary use, and correct grammar and mechanics. In this way, the scoring criteria is identical for each rater and they are implemented in the study in similar way thanks to the detailed descriptors in analytic rubric. Thereby it enhances consistency in scores because individual interpretation is eliminated (Hamp-Lyons, 1991; Weigle 2002; Johnson & Svingby, 2007). However, this method is often more time-consuming because each criterion must be considered separately (Perkins, 1983; Casanave, 2004).

Variance in Native and Non-native Scoring

Subjectivity has been one of the biggest challenges in writing assessments that are rooted in scoring variance in performance assessment. Likewise, the same challenge for writing assessments still endures because it also depends on the personal judgments of the raters. As Weigle (2002) indicates, there are two main themes of scoring variance. The first is “rater focus” and the other is “rater characteristics” (p. 70). In parallel with that, “rater focus” can stand for the differences in the rater’s decision-making processes including rubric use, their scoring behavior in terms of severity and consistency in rating whereas “rater characteristics” represent the rater’s background, culture, education, or native language. Many of them also address the distinctions in rating practices between native and non-native speakers as well as the rating habits of each group and they conclude varied but important results (Rao & Liu, 2020, Kim & di Gennaro, 2012, O’loughlin, 1992, Gonzalez Quintero, 2017; Sakyi, 2000; Shi, 2001 Lee, 2009; Santos, 1988; Hyland & Anan, 2006; Kobayashi, 1992).

Native speaker of a language is a person who acquires a particular language as his/her first language, from childhood (Longman, 2024). Native English speaker (NS) teachers and instructors, have an important place in ESL/EFL education for they actively take part in English language teaching all over the world. Not surprisingly, their approach to writing assessment and the differences has been an object of curiosity since they play a critical role in writing assessment as raters as well. In other words, as the natural speaker of English language the difference between their perceptions of a text has been expected to be different.

Furthermore, there is a division between two groups of research. One group claims that as natural inquirer of the English language, NS are attributed more reliable judgement of writing quality for they have broader knowledge of language. Thus, NS rater’s decisions are more accurate (Hughes & Lascaratou, 1982; Kobayashi, 1992; Takashima, 1987). The other group indicates that NNS rater’s assessment is no inferior to the NS as long as they are trained about writing assessment. However, sharing the native language with the test takers inevitably effects their judgement (Davies, 1983; Marefat & Heydari, 2016; Hyland & Anan, 2006, Hill, 1994).

It has been often investigated through many studies whether their assessment practice is somehow different in terms of decision-making process from NNS English teachers. (Carl, 1977; Rao & Liu, 2020, Cumming, 1990; Sakyi, 2000), rubric use (Gonzalez, 2017; Barkaoui, 2010), rater reliability and consistency (Shi, 2001; Lee, 2009) and scoring behavior (Kim & Gennaro, 2012; O’loughlin, 1992; Kobayashi, 1992). Some

has been conducted on the variability in scoring among raters with varying levels of experience and they revealed notable disparities in the application and prioritization of criteria (Pula & Huot, 1993), as well as in the severity and consistency of ratings (Weigle, 1998; McNamara 1996; Shin, 2010).

Decision-making process can be considered as an umbrella term referring to the factors such as rubric use and prioritization of rating criteria. Several empirical research has demonstrated the variance in the use of scoring methods and variance in scoring detected. Some of them manipulated analytical methods and pointed out the structural errors count more rather than the organizational or lexical errors while others resulted in an equality of importance for lexis and structure (McDaniel, 1985; Vaughan, 1991; Sakyi, 2000, Hinkel, 2003, Heaton, 1990; Hughes, 2003; Weigle et al., 2003). Additionally, in some research overall or global criteria associated with content and organization take the primacy in assessing the text (Hill, 1994; Song & Caruso, 1996; Lee, 2009). For holistic assessment, the results are also divergent. For instance, when the raters utilized the holistic method, they possibly felt freer, not constricted by definite criteria and they came up with other criteria that did not exist in the scoring rubric (Sakyi, 2000). Sometimes it is revealed by some studies that there is a significant difference between holistic and analytic scoring of the same text and the same rater (Russikoff, 1995) or some studies unveiled that the most determinative criteria, even the scores altered according to the rater groups (Santos, 1988; Song & Caruso, 1996; Brown, 1993; Shi, 2001; Lee, 2016).

These examinations focused on discerning categorical distinctions and the strategies that raters naturally developed based on their professional experience and training. The research not only analyzed behavioral differences between the two groups in terms of interpretation and judgmental strategies but also established a framework for raters' reading focus such as errors, topic, presentation of the ideas, and the scoring guide. Cumming and Sakyi (2002) conducted a multiple-steps-research to investigate the decision-making behavior of NS and NNS raters. In the first step, they transcribed the think-aloud sessions of NS and NNS rating processes. Then, they analyzed them to detect and group the categories they referred to during the sessions. For this, they marked the segments of the composition. Most of them included similar segments regardless of NS and NNS evaluation. Outlining rating categories was another attempt as shown in Table 5 (p. 77).

Table 5

Descriptive Framework of Decision-Making Behavior (Cumming et al. 2002)

Self-Monitoring Focus	Rhetorical and Ideational Focus	Language Focus
Read or interpret essay prompt	Interpret ambiguous or unclear phrases	Observe layout
Read or reread composition	Discern rhetorical structure	Classify errors into types
Envision personal situation	Summarize ideas or propositions	Edit phrases for interpretation
Scan whole composition	Assess reasoning, logic, or topic development	Assess quantity of total written production
Decide on macro strategy for reading and rating	Assess task completion	Assess comprehensibility
Consider own personal response or biases	Assess relevance	Consider gravity of errors
Define or revise own criteria	Assess coherence	Consider error frequency
Compare with other com positions or "anchors"	Assess interest, originality, or creativity	Observe layout
Summarize, distinguish, or tally judgments collectively	Identify redundancies	Classify errors into types
Articulate general impression	Assess text organization	Edit phrases for interpretation
Articulate or revise scoring	Assess style, register, or genre	Assess quantity of total written production
	Rate ideas or rhetoric	Assess comprehensibility

While there was no significant difference observed in NS and NNS ratings, an exception was noted in the NS raters' language use efficiency and creativity in interpreting essays. Conversely, NNS raters tended to provide shorter transcribed data. However, the research indicates a consistent trend where rater backgrounds, native language, and experience play influential roles. Evidently, NS raters exhibit greater concern and balance regarding rhetoric and language accuracy, whereas the NNS group tends to prioritize syntax, language use, or error classification and correction (Cumming et al. 2002). Additionally, the factors that affect the raters' judgements were investigated under two

domains: language-related and content-related (Sakyi, 2000). Song and Caruso's findings support these statements. In their study, they compared NS and NNS raters' use of holistic and analytic rubrics and identified a significant difference in holistic scores in NNS who awarded considerably higher grades than they did in analytic scoring possibly because they put more importance on overall organization rather than linguistic components (1996). Song and Caruso found that NS and NNS raters scored significantly different in holistic technique. While both native English speakers (NS) and non-native English speakers (NNS) used the rubric in a similar manner, there were differences in their scoring methods. Both NS and NNS agreed on the relative importance of content and organization (64%) and language use (36%) in the rubric, on average (1996). Considering analytical categories of the rubric, for both rater groups content and organization are the foremost categories for natives whereas non-natives considered all of the categories when rating (O'loughlin, 1992). Non-native speakers of English tend to prioritize grammar and sentence structure in their evaluation of writing, while NS tend to place greater emphasis on the coherence of ideas and argumentation, often adopting a more positive approach to scoring. Research indicates that essays tend to receive more positive feedback. Grades increase correlating with fewer comments on content and an increased focus on grammatical accuracy (Rao & Liu, 2020; Barkaoui, 2010).

Additionally, comparisons between Japanese and American raters reveal that despite significant differences in their evaluative processes, the final scores assigned by both groups are similar. Rater identity plays a significant role in writing evaluation, as the concerns, standards, and criteria employed by raters may vary considerably, leading to divergent directions in feedback provision regardless of the proximity of final scores (Connor-Linton, 1995). Early studies, such as Khalil's seminal research in 1985, shed light on the evaluation practices of Native American raters when assessing the utterances of non-native speakers. This study highlighted native speakers' heightened sensitivity to semantic errors compared to grammatical errors (Khalil, 1985). Furthermore, investigations involving academics from the United States and Japan demonstrate variations in approach to error correction, transition interpretation, and sentence construction, despite a similar frequency of error identification (Takashima, 1987). When comparing the evaluative perspectives of Japanese non-native speaker (NNS) raters with those of native speakers, it is evident that NNS raters tend to be more critical of grammar and sentence structure, despite their assessments not being deemed as entirely accurate. Conversely, native speakers tend to adopt a more positive stance towards the overall text, prioritizing clarity, organization, and lexis, areas that they seldom correct according to Kobayashi (1992) and Schaefer (2008).

Some studies have investigated the identification and treatment of errors by raters, with a particular focus on the significance attributed to various types of errors. This research has provided valuable insights into the examination of this issue. Notably, the findings revealed that non-native speakers (NNS) tend to adopt a strict approach with little tolerance for grammatical errors, including those related to word order and vocabulary usage. In contrast, native speakers (NS) tend to have a more flexible perspective, particularly regarding grammar, and prioritize intelligibility as a key criterion for assessment. (Hyland & Anan, 2006; Hughes & Lascaratou, 1982). Additionally, NS tend to give more positive feedback, although they give lower scores compared to NNS. Specifically, NS often offer positive comments on the clarity and readability of language. On the other hand, NNS tend to express more negative views, especially in the area of general knowledge, (Shi, 2001).

Given the complexities and variability inherent in writing performance evaluation, there is ongoing inquiry into whether NS raters possess more accurate and consistent judgment compared to NNS speaker raters, yielding a diverse array of research findings. To put it in a different way, It has been questioned if the NS judgement is more reliable that NNS judgement in writing assessment. Reliability is associated with variance in scoring attitude in term of severity and consistency in the scores.

Significant differences were found between NS and NNS raters' scoring behavior of rating severity, and rating consistency. Thus, those variations have been the focus of numerous investigations (Rao & Liu, 2020, Kim & di Gennaro, 2012, O'loughlin, 1992, Gonzalez Quintero, 2017, Ellis, 1986; Kobayashi, 1992; Casanave, 2004; Hamilton et al., 1993; Brown, 1994).

Consistency in scoring is directly related to rater reliability as mentioned above. As a result of the subjective nature of writing assessment and the differences in scoring behaviors, it is questioned that if the NS teachers may have a more reliable judgement and consistent ratings as the natural speakers of the language. Both writing assessment and feedback are inherently subjective processes that depend on individual opinion. For this reason, both of them may show a wide range of variety and this can effect scoring variability. Thus, there is research about the NS are able to detect the statement's suitability in a particular context as well as the accuracy of the statement (Ellis, 1986; Kobayashi, 1992; Casanave, 2004). Kobayashi mentions the dilemmas he experienced himself as a Japanese English speaker and an instructor (1992). He implies that as an English learner, he received various corrections and scores from different evaluators and he realized that some raters concerned only about grammar and mechanics while others considered meaning and organization as well. He further highlights that a second reader in comparison to his original wording even further revised certain expressions, initially corrected by a first

reader. As an English instructor, on the other hand, he sometimes had to leave sentences unmarked because he felt uncertain about the naturalness of the expressions, as a NNS English speaker, even though they had been written with correct grammar (Kobayashi, 1992). Depending on his own experience Kobayashi implied that native perspective in assessment is more dependable because the disparity of NS and NNS evaluation is rooted from intuition in that, NNS rate the texts with their explicit knowledge and that might be insufficient unless they are exposed the language enough to gain intuition in English language (1992).

Similarly, it is sincerely stated in some studies by the raters that as NNS ESL instructor at various levels, they have occasionally encountered moments in which they have been wholly at a loss as to whether to accept a particular sentence that is grammatically correct but potentially awkward to NS for other reasons. In most cases, they have just left the sentence uncorrected because of uncertainty and subsequently wondered to what extent nonnative ESL teachers are qualified to correct English compositions (Casanave, 2004). Shi also researched the statistical difference between NS and NNS scorings of 10 essays and noted that there is a significant difference between the reliability coefficients. NS are more reliable and consistent than NNS (Shi, 2001).

Lee (2009) investigated the same issue and came up with similar results of NS perform more reliable in scoring than Korean NNS. Moreover, he analyzes their reliability in different in linguistic features. For Korean raters, grammar and sentence structure has far lower reliability than NS evaluation. However, the global features such as content and organization reliability coefficient is closer for the two groups. These results imply that NS and NNS inter-rater reliabilities are significantly different. Rating competences are similar in global quality whereas structural scores vary more in NS evaluation, which may generate from the deficiency in NS language competence in terms of language use (Lee, 2009). Nevertheless, intra-rater reliability is firm and sound with both groups of raters and these results align with those of Eckes' study (Eckes, 2012; Lee, 2009).

Some studies result in little difference in consistency (Kim & Gennaro, 2012, Marefat & Heydari, 2016, Lynch & McNamara, 2008), and even fewer indicate that NNS are more consistent (Hill, 1994). However, the majority of the literature concludes the other way around. For most of them, soring reliability values of NS and NNS raters, and consistency depending on those values, are fairly close. It is important to note these values represent the scorings in total, which means there are some variations in rater's evaluation processes or their approaches to different criteria (Lynch & McNamara, 2008). In that, rater reliability could be achieved when the scores assigned by raters closely match each other numerically. However, it also implies that although the numerical scores may be similar, the

evaluation of specific text components and the raters' approaches to these components may differ, as illustrated in the subsequent table.

Table 6

Inter-rater reliability coefficients of NS and NNS raters (Lee, 2009)

Feature	NS Raters	NNS Raters
Content	.729	.664
Organization	.639	.673
Vocabulary	.556	.562
Sentence Structure	.630	.431
Grammar	.574	.297
Overall	.782	.738

The rater's scoring practices also affect reliability and rater consistency. To put another way, many studies have addressed the reliability and consistency difficulties by looking at how severely they score the texts, which influences their scoring method's dependability and consistency. As a result, notable distinctions were also seen in the scoring behaviors of NS and NNS raters with regard to the severity of ratings and their perceptions of the significance and complexity of rating criteria. Numerous studies have focused on those differences (Kobayashi, 1992; Casanave, 2004; Rao & Liu, 2020; Kim & di Gennaro, 2012; O'loughlin, 1992; Gonzalez Quintero, 2017). As for the research on scoring severity, the main results can be categorized into two main perspectives. One perspective suggests that NNS raters tend to give harsher ratings compared to NS 1985; Santos, 1988; Fayer & Karshinski, 1987; Kim & di Gennaro, 2012; O'loughlin, 1992; Rao & Liu, 2020; Hill, 1996,Ross, 1979; Song & Caruso, 1996). The second perspective propose that while NS are still the most severe raters, NNS raters, on average, tend to be harsher, possibly due to a higher bias when both rater and examinee are from the same nationality (Kim & di Gennaro, 2012; O'loughlin, 1992; Lee, 2009).

According to some studies, the NS and NNS rate scores differ not only in terms of the general severity of scoring, but also in terms of how they score differently on specific assessment criteria (Eckes, 2008; Lee, 2009; Kondo-Brown, 2002; Shi, 2001; Schaefer, 2008; Shi, 2001; Weigle, 1998, Choi, 2002). The findings show that when it comes to grammar accuracy, NNS raters are more critical than NS raters (Eckes, 2008; Kondo-Brown, 2002; Schaefer, 2008; Hyland & Anan, 2016). They might be more severe when it comes to organization (Shi, 2001) or coherence and cohesion (Hill, 1994). However, NS

raters are typically more critical of content and structure and have a more positive attitude toward language and content, giving these factors precedence over linguistic aspects (Lee, 2016; O'loughlin, 1992, Hill, 1994, Shi, 2001). However, there are some research indicating reversely that NS are more considerate about grammar and NNS are about Vocabulary (Carl, 1977, Kobayashi, 1992; Connor-Linton, 1995).

Research has also been conducted in different countries, with different groups of NNS raters. In Iran, for instance, some scholars wondered about the variation of ratings between NS and Iranian raters. Marefat and Heydari observed the raters decision-making process through think-aloud sessions and they came up with the idea that there is significant difference in their opinions of what makes a good writing and in terms of severity. They also found that NNS were considerably more severe than the NS. Apart from its peers, and being inspired by Lee (2009) they also investigated raters' notions on the difficulty and importance of the criteria. They had the raters rank the categories from the most difficult to the least as well as the importance. Iranians ranked grammar as the easiest to rate and the most significant whereas NSs chose organization. On the contrary, Iranians classified organization as the least important and the most difficult to rate. Therefore, the study revealed a certain contradiction of ideas of his (2009).

Similarly, Rao and Liu conducted a study in China aiming to reveal to what extent to which NS and NNSs' scores vary when they assess Chinese students' writing samples of an English major department. The results show there is significant difference between the two groups in terms of scoring behavior (2020). These results suggest that NS, having never undergone an English as a L2 learning process and using English as their native language, might be indifferent for their mistakes in language use and as English their native language, they might consider a wider spectrum of criteria. On the other hand, NNS raters (all Chinese) might be less informed about western culture so do not expect as much as the natives do. (Rao & Liu, 2020).

As for Koreans, it is found that they performed a more severe attitude against grammar and sentence structure than their NS counterparts did although both groups perceived content the most difficult and important criteria for a well-written paragraph. In other words, NNS cared more about linguistic features when both groups came to terms on the significance of the global terms. However, it is implied in the study that Korean's low level of inter-rater reliability may indicate unsteadiness in their eligibility in writing assessment (Lee, 2009). In another study regarding the scoring behavior of experienced and inexperienced NS raters in Korea, it is implied that natives exhibit a stricter attitude towards content in comparison with the NNS who graded the linguistic criteria more

thoroughly (Lee, 2016). In parallel with that, Japanese raters' severity in scoring and they display personal strategies when they response to the text (Schaefer, 2008).

In contrast with those studies that imply NNS are inferior in writing assessment, there are others to claim little difference exists depending on the mother tongue about reliability (Hill, 1994; Cumming et.al, 2002; Song & Caruso, 1996, Brown, 1996). Hill, for example, is one of the researchers who question the suitability of NNS raters. The research was conducted between Australian and Indonesian raters ended up with proximate statistics and that implied NNS are as qualified and reliable as NS raters (Hill, 1994) are. Statistics did not point out any evidence to prove that any group is more suitable. Other than that, it is, similar to the previous research, it is estimated that both groups are as consistent in evaluations. However, contrarily to most of the studies, NS were hasher in this one (Hill, 1994). Nevertheless, the NNS in this research were not prone to award high marks probably due to their lack of confidence in language proficiency and their belief in NS superiority in language. These results align with Hyland & Anan's study with Japanese NNS. In this study, were referred as "error hunters" performing remarkable severity in grammar and they adopted grammar error correction role while assessing. In addition, NNS are implied to lack of confidence because of limited opportunity to expose English language to gain proficiency (2006). Likewise, Brown found no significant correlation between NS and NNS speakers regarding harshness in his study where he compared NS and NNS ratings in an English speaking test for tour guides in Japan and it cannot be inferred that NS raters are more suitable in performance evaluating with this results (1995).

Conclusion

Writing assessment has always been a complicated issue in language learning. It is a complex skill since writers go through a bunch of complex cognitive processes to be able to write a well-organized text. They need to master some linguistic, structural, and cultural concepts. In addition, the writer's background and characteristics influence the text, making it a personal, unique production. Writing assessment requires assessing the performance of an individual's work and the rater's background, such as their experience, native language, education, or personality are involved in the assessment process. Thus, scores vary depending on the rater's background and approach, which causes subjective results. Although subjectivity in this issue looks inevitable to some extent, it is to ensure its reliability because one's writing competency can play a crucial role in global or local language tests whose results are regarded for employment, or obtaining education or scholarship opportunities. Therefore, this issue has not slipped past the notice of scholars. The studies

have shown that subjectivity and reliability problems in writing assessments originate from scoring variance in the decision-making procedure, rater consistency, and scoring behavior. This study aims to shed light on the scoring variance of NS and NNS raters, and in the next chapter, the methods can be found.

Chapter 3

Methodology

Introduction

The current study aims to analyze the NS and NNS writing assessment approaches, identifying the variances with their reasons. In this part, the systematics of the research design and the rationale behind will be elaborately discussed along with other details such as participant profile, research setting, data collection steps, and instruments.

The investigation mainly concerned decision-making behaviors and scoring behaviors of the raters. Decision-making behaviors are the strategies the raters adopted while assessing a text whereas scoring behaviors are related to severity and their ways of criteria use. In addition, the differences between the raters' approach to the holistic and analytic assessment are able to be inferred since the study includes an implementation of each scale.

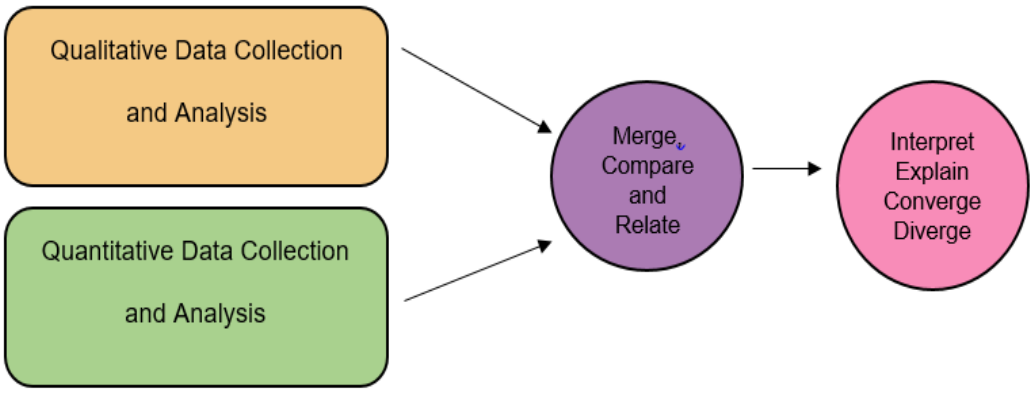
Type of Research

The purpose of this study is to investigate the scoring variance of NS and NNS raters' subjective judgment in writing assessments in terms of their decision-making behavior, scoring behavior, and use of the scoring method. The study includes both numeric data (scorings of the writing samples) and verbal data (interpretations of writing samples and interviews). For this reason, this study employed a mixed method where both quantitative and qualitative methods were used. The mixed research method refers to the hybrid "combination of quantitative and qualitative methods within the same study" (Dörnyei, f It is an advantageous method since the two research methods can be intertwined in the way that they compensate for each other's deficiencies. For example, qualitative methods can be criticized for being too "context-driven" and "unrepresentative" whereas quantitative methods are implied to be too far from the effect of life and circumstances and "decontextualized". However, these weaknesses can be eliminated in the case of converging the qualitative and quantitative methods in one single study by achieving a more thorough understanding and approaching the matter from different perspectives. (Sandelowski, 2003; Dörnyei, 2007). This convergence of different methods, also called 'triangulation', enables us to verify the argumentation of the data (Dörnyei, 2007). As for this study, 'convergent parallel mixed method design' was adopted in particular (Creswell, 2021). In this method, the researcher collects and analyzes qualitative and quantitative data

separately before they integrate them and withdraw inferences, meaningful divergence, and convergences. In this way, the researcher can look at the findings from varied angles (Creswell, 2021). For this particular research, quantitative data were collected via a Google Forms questionnaire including questions about the participants' scorings of some sample writings of the students in a preparatory class of a state university in Turkey. In addition, qualitative data depend on think-aloud sessions where the participants' scoring process can be observed. This part is followed by an interview session with semi-structured questions to understand their approach more comprehensively. The focus of the study is the qualitative data, which is composed of, think-aloud sessions and interviews.

Figure 7

Convergent Parallel Mixed Method Design - QUAL> quan (adapted from Creswell, 2021)



Participants

This study was conducted at a preparatory class of a state university in Turkey. 20 instructors, 10 NSs and 10 NNSs, participated in the study as writing raters. The participants were chosen utilizing purposeful sampling, a qualitative research sampling technique. This method involves deliberately selecting participants based on their expertise to yield insightful data on a specific topic. In this study, participants were selected from both NS (Native Speaker) and NNS (Non-Native Speaker) instructors who teach and assess writing at various universities in Turkey. Their expertise directly aligns with the research topic, ensuring the relevance and depth of the data collected (Bernard 2002, Creswell, 2021). Three of the NS are employed by that state university in addition to other institutions they also work for. The seven remaining NS work for other state and private universities in Turkey. All of them have currently been teaching preparatory classes. Their experience

ranges up to 20 years and education levels vary as graduate, MA, and PhD. All of the NNS raters were Turkish and nine of the NNS raters worked for the aforesaid state university. The participants varied in education and experience. Out of the total twenty participants, five were male, and fifteen were female. Most of them had 6-10 years of experience. Only 5 of them had 20 and more years of experience and all of them were NS raters. In terms of their education, most of the raters, 14 of them had masters degree and 2 NS had PhD degree. All of them were ELT or equivalent training. In other words, the NS and NNS participants experience and education are generally in parallel with each other. However, these aspects are not in the scope of this particular study although they are considered by many others in the literature. They were tasked with rating 12 achievement exam writings from elementary and intermediate level students at the university. Ratings were conducted using both holistic impression scoring and analytic scoring with a rubric. Though the students demographics are not a considered as a variable for this study, they came from different cities of Turkey with ranging levels of English. Their ages changed from 17 to 20. Their levels are diagnosed by a placement exam at the beginning of the semester and they complete all year in the same group of students in the same class. Consent forms were obtained from students and raters for their participation in the study (Appendix E and F). The institution was duly informed and granted approval for sharing the writing samples. Students undergo four hours of skills classes, covering listening, speaking, reading, and writing. Writing practice is integrated into the curriculum through assigned paragraphs or essays tailored to their proficiency level and current unit. Assessment of writing skills occurs during achievement and proficiency exams held throughout the semester. The students should collect at least 60 points in 4 achievement exams and then they should get at least 60 points out of 100 in the proficiency exam. Texts are analytically evaluated by two instructors. The raters consider the criteria consisting of content, organization, grammar, vocabulary and mechanics. They can rate between 5 and 0 points. If there is a discrepancy of three or more points in the writing scores, a third rater assesses the texts. Writing samples retrieved for this study were equally distributed to NS and NNSs by depending on the students' proficiency levels by the researcher.

Data Collection

In this study, both qualitative and quantitative data were collected to investigate [insert purpose or aim of the study]. All data instruments were transcribed on Google Forms to facilitate accessibility for participants during remote sessions and for practical use in data analysis. The questionnaire comprised three parts: the first part gathered information about

participants' background, gender, experience, and education, along with questions to assess their understanding of key points. Further details about the questionnaire can be found in the 'Instruments' section later in this chapter.

Following the completion of the questionnaire, participants proceeded to the second part, where they evaluated writing samples during think-aloud sessions. Prior to rating, participants watched a YouTube video demonstrating a think-aloud session, considering that some participants may have been inexperienced in such sessions to ensure reliability. Throughout the sessions, the researcher provided guidance and addressed any questions that arose. Sessions were conducted via Zoom or face-to-face, accommodating participants' preferences and locations. To maintain fairness, each rater communicated in their NS language. Turkish transcriptions were translated into English post-collection. Participants provided written and spoken consent for voice recordings, which were transcribed via Microsoft 365 later. All transcriptions were made using Microsoft 365. After that, the researcher reviewed the transcriptions. Sessions lasted approximately ninety minutes. "The participants were asked to rate 12 writing samples. The researcher transcribed samples carefully, maintaining the exact wording and organization of the original student writing. The writing samples were divided two groups in six. In each group, there were three elementary, three intermediate level samples.

Following the questionnaire, the participants were required to rate the first group of texts in six through impression rating, a holistic scoring method; and the other using a rubric; an analytic scoring method. The assessment process is administrated individually. The transcriptions provided qualitative data, which was analyzed to identify the rater's scoring steps to find out their decision-making behaviors. The scorings formed the quantitative data that enlightened the issues of inter-rater and intra-rater reliability, rater consistency, and severity in scoring. The rubric (Appendix H) was adapted by the researcher (Council of Europe, 2009; CEFR, 2002; Jacobs et al., 1981; Weir, 1990; Gonzalez, 2017) and revised by two experts. Both of these experts were professionals in writing assessment, who work in the preparation classes in two different state universities in Turkey. They were both educated in the field and one of them is the head of the writing center at her institution. The rubric was approved in terms of its convenience by these experts. A detailed description of the rubric is presented in the 'Instruments' section as well (Appendix H). For the last part, the participants were interviewed about their approach to writing assessment. The researcher prepared the semi-structured interview. The main aim was to reach a more comprehensive understanding of their thoughts on the process and explain themselves.

Instruments

In this study, the data were collected through three techniques that are the background information questionnaire, think-aloud-session scoring of the writing samples, and semi-structured interviews. The study involved fewer than 30 participants, categorizing it as non-parametric. Initially, they were given the consent forms (Appendix E) indicating their approval of the study. They were informed about the study by the researcher. Following the display of a video demonstrating the think-aloud process (Think-aloud Protocol by Ozgur Sahin, 2016) the participants completed the questionnaire which gathered information on their background knowledge and perceptions of writing assessment through close-ended questions. The subsequent task involved scoring 12 writing samples and verbally evaluating them. These samples, derived from the first achievement exam of elementary and intermediate students, were grouped equally into six based on the grades awarded by two raters at the institution. Finally, participants engaged in semi-structured interviews aimed at gaining insight into their approach to writing assessment.

Table 7

Writing Tasks

Lower level (Elementary)	Higher level (Intermediate)
Task: Write a descriptive paragraph about your dream hotel resort. Write a topic sentence and three supporting ideas in you paragraph. (100-200 words).	**Task:** Read the statement below. Write an opinion paragraph according to the prompts (150-200 words). "Some people believe that studying together in a group is more beneficial than studying individually." To what extent do you agree or disagree? Write one paragraph including; a topic sentence two supporting ideas for one opinion and give an example for each opinion two supporting ideas for the contrasting opinion and give an example for each supporting idea a conclusion Use appropriate linkers, tenses, transitions between paragraphs.

Instrument 1 – Questionnaire

The first instrument utilized in the current study is the background questionnaire, developed by the researcher. This questionnaire, housed in Google Forms for practicality in reaching distant participants and data analysis, underwent content validity revision by two experts. An example of the from can be found in Appendix G. Their feedback helped identify typos and improve clarity. Comprising 12 close-ended questions, the questionnaire delved into participants' background information such as gender, experience, and education, primarily in 'yes-no' or short-answer format. Additionally, participants were questioned about their perceptions regarding scoring reliability, the importance of writing criteria, and challenges encountered in rating certain criteria. Finally, they were asked about their scoring behaviors in general. These responses were collected with the aim of comparing them with participants' scoring behavior during think-aloud sessions and their interview responses, facilitating a comprehensive understanding of their assessment practices.

Instrument 2 – Think-aloud Sessions

The second instrument utilized for the data collection part is think-aloud sessions, a technique requiring the participant to verbalize everything they think when they complete a task. Afterward, the data are transcribed and analyzed. As for this study, participants were required to assess the 12 writing samples aloud while articulating their interpretations of the texts and explaining the rationale behind their awarded scores. This method was deemed the most suitable for our research objectives as it allows for the interrogation of raters' cognitive processes, shedding light on their decision-making rationale. In furtherance, the literature indicates that think-aloud session is an effective method to perceive “unobservable cognitive processes” (Barkaoui, 2008). Moreover, Kasper (1987) suggests that such sessions offer insights into informants' overarching approach to a task, the decision-making levels involved, and the factors influencing their judgments. Similarly, this study aims at detecting the characteristics that strengthen and develop a written piece in the rater’s opinion. In other words, it searches for the most important criteria that make writing beautiful from their perspectives. Think-aloud sessions are also quite functional for this aim as Huot (1993) states that we can pursue “what the participants are really concerned about” through think-aloud sessions. It is not “filtered”, and it does not include “generalizations” which are unlikely to be provided in other self-report methods since the researcher can elicit the “actual behavior” of the participant, not merely what they say (Connor-Linton, 1995; Huot, 1993). Thus, think-aloud sessions emerge as the most convenient research method for it allows the researcher to find out hidden nuances and thorough examination of the subject matter.

Two scoring methods were implemented in the study in a way that their advantages and drawbacks complete each other. For example, impression scoring is a sub-method of the holistic scoring. No holistic rubrics is necessary and assessment depends on the raters' innate judgement and personal knowledge. Impression reading is efficacious for gathering the readers' original thoughts on writing assessment as this very study targets since there is no use of a rubric to direct the raters' judgement. They employ their personal criteria according to their individual impression (Baxter, 1997; Hughes, 2003).

The researcher modified the rubric and had it revised by two experts after it was first adopted (Council of Europe, 2009; CEFR, 2002; Jacobs et al., 1981; Weir, 1990; Gonzalez, 2017). It included five criteria that are content, organization, and use of language, vocabulary and grammar. The description of the scope of the criteria was as follows:

- **Content**: It regards the content quality of the paragraph. It is expected that

the required conditions in the question stem are met. If a certain conjunction, grammar, topic, word limit, etc. are covered in the question stem, the answer will be evaluated according to whether these are met or not.

- **Organization**: It regards the logical and meaningful organization of the paragraph. The semantic integrity and fluidity of the writing answer will be evaluated in terms of word-subject repetition and appropriate conjunction use.

- **Use of Language:** It regards language accuracy of the paragraph. The student's use of grammar rules he/she has learned within the scope of the course curriculum will be evaluated.

- **Vocabulary:** The student's use of the words and word types he/she has learned within the scope of the course curriculum will be evaluated.

- **Mechanics**: It regards the rules that the student has learned within the scope of the course curriculum, use of conjunctions and capital letters, punctuation marks, spelling mistakes, legibility, and paragraph formation.

Subsequently, this method provided some clues about the reliability and utilization of holistic assessment. Whether there is a significant difference between NS and NNSs raters' holistic assessment was also hinted. In the second part of the writing samples, the raters' made use of an analytic rubric. The main purpose here was to identify to which criteria is considered and to what extend it is highlighted by the rater. Additionally, the possible variations in rubric use and reliability of rubric use was observed. Thus, the data collected allowed us to make a comparison between two group of rater's holistic

(impression) and analytic rating processes. Finally, as the study included elementary and intermediate level writings, the difference in raters' scoring behavior can be observed according to the stages.

Instrument 3 semi-structured interviews

Interviews were the last method used in the study to collect data. The researcher's purpose was to gain a deeper understanding of the participants in writing assessment for this method serves to discern the insights of their approach (Dörnyei, 2007). They mentioned their personal opinions independently. Every remark and expression was recorded carefully not to evade any details. From time to time, the researcher contributed to the participants' comments or asked further questions both to foster them to share more ideas and to understand their point deeply. The interview part functioned as a follow up part of the think-aloud session. They were prepared by the researcher herself considering the literature and assumed results of the study. The target of the interview questions was to uncover the hidden implications in the rater's approach acquired from think-aloud sessions. All participants were interviewed with the questions below:

1. *Do you enjoy assessing writing?*
2. *Do you encounter any challenges or difficulties when assessing writing?*
3. *Which scoring method is more useful in your opinion? Why?*
4. *Do you think writing are assessed fairly today?*
5. *Do you think it's more reliable when the writings are assessed by native raters?*

Table 8

Data Collection Instruments

Design	Research Questions	Data Collection Instrument
Quantitative Data	1. To what extent do Native and non-native raters diverge in their assessment of identical paragraphs?	Scoring results of NSs and NNSs
Qualitative Data	2. How do Native and non-native raters consider various criteria when evaluating writings?	Think-Aloud Session Interview
Participants	10 NS and 10 NNS whose native language was Turkish, attended the study. All of the participants have worked in the preparatory classes of several state or private universities in Turkey and all of them were educated in English language teaching (graduate, MA, and PhD).	

Data Analysis

The purpose of the current study is to reveal NS and NNS writing assessment methods, noting differences and providing explanations for them. Both qualitative and quantitative data were collected. Qualitative data included clues about the raters decision-making behaviors which refer to the strategies they use while rating a paragraph. This investigation uses the 35 decision-making behaviors identified in Cumming and Sakyi's study (2002). The researcher transcribed the qualitative data from the interviews and the think aloud sessions analyzed the data herself. While analyzing the transcriptions, the opinions of two experts with PhDs in writing assessment were consulted. The qualitative data was transcribed through Microsoft 365 and controlled by the researcher before it was analyzed through thematic analysis method. The quantitative data depended on the scorings of the writings. It was collected through Google Forms and utilized for investigating descriptive analysis The research's data were moved into a computer environment, where they were arranged using the Microsoft Excel package and analyzed using the SPSS (Statistical Package for Social Sciences) 29.0 tool. The statistics shed light on the numerical

variance among the consistency of a rater's scorings (inter-rater reliability), and their scoring behavior on scoring severity. In addition, significant differences and variations in holistic and analytic scoring and the rater's use of rubric criteria can be inferred from the numerical data, which was used to compare with the qualitative data.

Conclusion

This study employs a mixed-method approach, utilizing both qualitative and quantitative data. Twelve paragraphs written by university preparatory class students were scored by groups of Native Speaker (NS) and Non-Native Speaker (NNS) raters. The scores provided the quantitative data, while think-aloud sessions generated the qualitative insights. The data were then analyzed and synthesized to draw conclusions

Chapter 4
Findings and Discussion

Findings

Introduction

In this study, the data was collected through three methods which depend on the NS and NNSs' scorings of 12 paragraphs, think aloud sessions of the rating process of the two groups of participants and the interviews with them. In this section, the statistics of the scorings of the paragraphs and the thematic analysis of the think aloud sessions and the interviews are presented. This study is non-parametric due to the number of the participants is under 30. Since the native language effect on ESL/EFL writing assessment is researched, the participants other characteristics and background were ignored. The data obtained as a result of the research were transferred to the computer environment and organized with Microsoft Excel package program and then analyzed with SPSS (Statistical Package for Social Sciences) 29.0 package program. Categorical data were shown with frequency and percentage values. Numerical data were analyzed using a nonparametric test since the number of individuals in both groups was less than 30 and shown as mean, standard deviation, minimum and maximum values. Differences between native Turkish and English speakers were analyzed using the Mann-Whitney U Test. Statistical significance level was accepted as $p<0.05$. The qualitative data collected through the think-aloud sessions and interviews, which lasted 90 minutes in average, were analyzed by the researcher. The abovementioned paragraphs were written by elementary level and 6 Intermediate level students. They were named from Paragraph 1 to paragraph 12. The first group of six paragraphs included 3 elementary and 3 intermediate paragraphs and they were scored via impression scoring method whereas the second group of six was scored by using an analytic rubric. The paragraphs are demonstrated in Table 9 below. All paragraphs were scored by 10 NS and 10 NNS raters. The findings is presented in two part according to the research questions concerning the quantitative and qualitative information.

Table 9

The grouping of and scoring of the paragraphs

	TOTAL SCORES OF THE PARAGRAPHS											
	IMPRESSION SCORING						ANALYTIC SCORING					
	Elementary			Intermediate			Elementary			Intermediate		
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
NS 1	12	16	20	5	20	15	11	14	18	13	15	18
NS 2	15	10	19	12	17	19	14	15	12	17	19	18
NS 3	19	17	23	17	25	20	18	24	20	21	18	20
NS 4	20	18	23	18	20	18	19	18	27	17	22	20
NS 5	13	18	20	16	22	22	20	21	17	24	23	21
NS 6	19	15	23	12	23	24	14	12	15	18	29	19
NS 7	18	12	22	10	22	18	16	16	19	17	21	18
NS 8	12	14	16	15	18	20	13	17	17	17	20	19
NS 9	19	13	20	5	20	18	18	18	20	19	20	18
NS 10	18	8	24	8	21	5	20	16	15	15	14	17
NNS 1	15	12	20	7	21	13	12	9	17	15	20	15
NNS 2	17	16	20	15	20	19	16	17	19	18	20	17
NNS 3	17	7	25	10	18	21	18	17	22	16	20	19
NNS 4	18	13	21	13	20	20	12	12	19	13	18	11
NNS 5	11	14	20	7	21	15	13	18	23	15	18	17
NNS 6	18	12	21	10	18	15	19	14	18	15	17	14
NNS 7	18	19	23	16	19	20	13	18	19	21	25	18
NNS 8	21	14	23	12	21	22	18	16	17	19	22	21
NNS 9	19	17	23	14	21	19	15	11	19	11	18	12
NNS 10	16	12	24	12	20	22	20	16	15	15	14	17

Findings for the Research Questions:

Research Question 1: To what extent do Native and non-native raters diverge in their assessment of identical paragraphs? This research question aims to reveal any possible scoring variance in NS and NNS raters' judgement in terms of rater consistency and severity. To answer this question, NS and NNS raters scored 12 writings both through analytic scoring method with a rubric and via impression scoring, which depends on the raters' free judgement and requires no scale or rubric. First 6 paragraph were scored via

impression scoring (from Paragraph 1 to Paragraph 6) while the last six were scored with an analytic rubric (from paragraph 7 to Paragraph 12).

The main objective of including impression scoring to this study is to be able to detect the participants decision-making behaviors and their personal strategy of assessing writing without a guide such as a rubric. In other words, through impression scoring, it is intended to uncover their approach without any intervention. Regarding grading, impression scoring method is considered not as reliable as the analytic method for the raters' personal stance. However, unlike the expectations, the results of impression scorings of NS and NNS raters ended up with quite consistent values (Table 10).

Table 10

First 6 writings' Median, Minimum and Maximum Scores of Impression Scoring

Impression Scoring		x	SS	Median	min	max	Z	p
Paragraph 1	NNS	17,00	2,67	17,50	11,00	21,00	-0,153	0,878
	NS	16,50	3,17	18,00	12,00	20,00		
Paragraph 2	NNS	13,60	3,31	13,50	7,00	19,00	-0,457	0,648
	NS	14,10	3,38	14,50	8,00	18,00		
Paragraph 3	NNS	22,00	1,83	22,00	20,00	25,00	-1,136	0,256
	NS	20,90	2,33	21,00	16,00	23,00		
Paragraph 4	NNS	11,60	3,10	12,00	7,00	16,00	-0,967	0,334
	NS	11,80	4,76	12,00	5,00	18,00		
Paragraph 5	NNS	19,90	1,20	20,00	18,00	21,00	-0,305	0,760
	NS	20,80	2,35	20,50	17,00	25,00		
Paragraph 6	NNS	18,	3,17	19,50	13,00	22,00	-1,991	0,046*
	NS	17,90	5,15	18,50	5,00	24,00		

Mann-Whitney U Test, p<0,05

Analyzing the table reveals that there is no statistically significant difference in the ratings of paragraphs 1, 2, 3, 4, and 5 between NNS and NS. (p>0.05). The scores' difference from the mean is indicated by the standard deviation value. When the table is analyzed, It can be seen that in all of the NNS gave more similar scores, while NS gave more different scores. These findings indicate that there is considerable overlap in the score distribution for these paragraphs. Standard deviation value also indicates that paragraph 3 and paragraph 5 are assessed more consistently by both of the groups. These two paragraphs have the highest minimum scores in the group. It can also be concluded that paragraphs 1, 3 and 5 are higher in quality. On the other hand, paragraphs 2, 4 and 6 are the highest standard deviations values meaning that these are the least consistently scored

paragraphs by both of the groups. In that, both NS and NNS raters had varied judgements and scorings are the most diverted in these three paragraphs. Considering the minimum and maximum scores, it can be inferred that they are weaker and poorly-written paragraphs in comparison with paragraphs 1, 3 and 5.

These assessments lead to the conclusion that the first five paragraphs are often assessed consistently by the both groups of raters. In terms of severity, NS and NNS groups' ratings showed parallelism as evidenced by the minimum and maximum scores for the first 5 paragraphs are fairly close to each other, indicating that, there is no significant difference in terms of scoring behavior between the two groups. To put it differently, both of the groups score almost equally severe or lenient in their scorings (Table 10).

The assessment results for paragraph 6 were the only ones that revealed a statistically significant difference ($p<0.05$) between NS and NNS. NS scored 17.90±5.15, but NNS scored 18.60±3.17 on the Paragraph 6 examination. NNS, the scoring was more lenient. NS scored between a minimum of 5 and a maximum of 24, while NNS scored between a minimum of 13 and a maximum of 22. For NS, the score range is higher which means that the NS raters' scores and interpretations vary more than the NNS. One of the NS assigned the harshest score whereas another NS graded the most lenient score that means that NS scored less consistently compared to the NNS raters because the score range is wider. To put it differently raters did not come to terms on its quality and made different interpretations. Their positive, negative and neutral interpretations showed variation, as shown in the table 11.

Table 11

NS Raters' Opinions about the Paragraph 6

Positive	Negative	Neutral
"The content is well-developed and it keeps communication. Some problems with word forms and spelling. I give a bit higher Because their command of the vocabulary is better" (NS 1)	"I'm going to be honest. That one was chaos. It was really, really hard to understand" (NS 7)	"That' a tough one because It's not terrible writing, it's just not fulfilling the assignments. Well, I think that's maybe a slight misunderstanding of the task." (NS 4)
"It gives all necessary parts and its easy to follow." (NS 2)	"I'm confused. He tries to summarize his ideas but they are not communicated well" (NS 10)	"There some mistakes but that is OK: We work in ESL. We're expecting the writings to be very good, but not perfect." (NS 6)

In the table above (table 11) NS raters' positive, negative and neutral opinions about paragraph 6 can be seen. These are chosen from 10 opinions of the raters to represent the opinion variation about the identical paragraph. NS 1 and NS 2 found the paragraph adequately developed to do the task and to facilitate communication. NS 4 and NS 6 expressed both positive and negative aspects of the paragraph 6 NS 6 and NS 7 made the harshest comments on it. It can be concluded from this table that raters can evaluate and assess the same paragraph very differently. Their opinions may change in terms of severity. When some of them make lenient comments, others can tend to give harsher opinions and to cut off points. This situation also effects rater consistency negatively.

Table 12

NNS Raters' Opinions about the Paragraph 6

Positive	Negative	Neutral
"The student can get the message across and there is kind of an organization" (NNS 1)	"The paragraph is not well-organized. There are too many mistakes and the vocabulary is insufficient" (NNS 7)	"I can understand what he means but there are some grammar mistakes that effect meaning " (NNS 4)
"It is very easy to understand and follow" (NNS 3)	"Supporting ideas and the content are not developed. I don't like the students ideas" (NNS 3)	"Here, content and organization are good. It answers the question but the topic sentence is copy paste." (NS 6)

Considering the table above, the raters' opinions about the quality of a students' writing performance may vary depending on the criteria that are important to them. The raters scored the paragraph through impression scoring which is a holistic method. In this method, the raters rate more freely and their personal opinions are influential.

According to the information in the tables 11 and table 12 some NS made positive and more tolerant comments on the paragraph 6 while there were also negative and neutral opinions. The same case is also true for the NNS raters. From this table, it can be understood that raters' thoughts on a paragraph and the student performance may differ and it reflects to the scorings. As it can also be seen in the tables, 11 and 12 some comments may be more severe or lenient. When one rater describes the paragraph well developed and communicative, another one may find the identical paragraph "chaotic". Therefore, both the NS and NNS raters' scoring behaviors in terms of severity an leniency do reflect both to their scorings and also to their opinions even though there is no significant difference between the groups in paragraph 6.

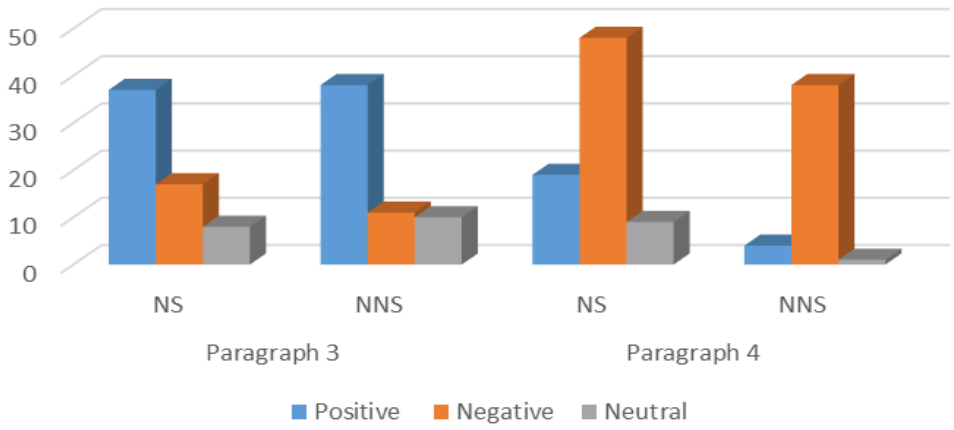
Table 10 demonstrates scoring variance in the impression scores for the paragraphs by the NS and NNS raters. As the means suggest, in paragraph 3, the standard deviations for NS and NNS ratings were under 5 ($p<0.05$), which means that the range of scores was not high and there is no significant difference between the groups in terms of their ratings. This refers to high rater consistency in both the NS and NNS groups' ratings. As for

paragraph 4, the standard deviation was much higher than for paragraph 3, although it also fell below five ($p<0.05$) and did not indicate any significant difference. It means that the range of scores awarded to paragraph 4 is a lot wider than paragraph 3. Moreover, according to the ratings, both the NS and NNS raters implied that paragraph 3 was remarkably more qualified than paragraph 4. Therefore, It can be concluded that the more qualified a paragraph is, the narrower the range of scores, which points to higher rater consistency. It is clear that both the NS and NNS exhibit a similar pattern of rater consistency, as there is not a significant difference in their scorings. Their scorings fluctuate in parallel with each other according to the quality of the writing.

In addition to rater consistency, the NS and NNS raters' scoring behavior can be identified. As it is demonstrated in the table, paragraph 3 received the highest scores in average whereas paragraph 4 received the lowest scores indicating that paragraph 3 is better developed than paragraph 4. In the table 10, the lowest score for the paragraph 4 was given by the NS raters, which also illustrated that NS raters were slightly more severe in their scoring behavior than NNS raters. As for the paragraph 3, the NNS are slightly more lenient in their scoring behavior. Overall, the severity in the scorings of both of these groups were not significantly different from each other in impression scoring. However, the rater's comments about paragraph 3 and paragraph 4 diverge in accordance with the quality of the paragraph. The table 13 below exhibits the number of the positive, negative and neutral comments on the paragraphs 3 and 4 made by the NS and NNS raters in the impression scoring part. Most of the NS and NNS raters made more comments that are positive on the paragraph 3 and found it better written. Most of the NS and NNS raters made more negative comments about the paragraph 4 and they evaluated it as a poor paragraph. Namely, most of the NS and NNS raters' comments aligned with each other's' assessments of the strongest and the weakest paragraphs of the first scoring part.

Table 13

Differences in the comments between Paragraph 3 and Paragraph 4



This increases the rater consistency in the assessment of these paragraphs. Nevertheless, in this table, it can be seen that there is a considerable difference between the two groups of raters regarding the numbers of positive comments on the paragraph 4. NS raters made comments that are more positive on the paragraph 4. This numbers indicated that the NS raters tended to articulate positive aspects of a paragraph compared to the NNS speakers. NNS raters made just a few positive comments on it. Some examples from the positive comments on NS and NNS raters are listed in the table 14 below. From their statements, it can be seen that the NS considered and articulated positive aspects more than the NNS raters articulate.

Table 14

Positive Comments on the Paragraph 4

NS	NNS
"They are trying to do the task. But their English is not competent enough to express the ideas they want to express clearly" (NS 1)	"We have a topic sentence. Yeah, we got linkers. Yes, we have something like a conclusion. Actually, this student knows how to write a paragraph. " (NNS 1)
"In terms of organization it's kind of doing what's asked for. In terms of the content. Yeah, I mean it's it is sort of answering the question, trying to use different words" (NS 4)	"Beneficial is a good vocabulary choice" (NNS 2)
"Surprisingly, this person has good ideas." (NS 5)	
"The style is a little different than what we usually assess but it's cool to see kind of a difference." (NS 9)	

While impression scoring was initially included to observe raters' personal strategies and approaches to writing assessment, the scores yielded valuable insights into scoring and interpretation variances during think-aloud sessions. However, it is important to note that assessing writing without a rubric is not as reliable as analytic scoring. Without a rubric, it is challenging to obtain information about how raters perceive specific criteria. In contrast, analytic assessment, which employs a rubric, breaks down scoring into specific criteria. This allows for analysis of discrepancies in scoring and raters' approaches to these criteria. Therefore, the second part of the study included analytic scoring to gather information on how raters approach analytical assessment and use rubrics. For this purpose, the second part analyzed writings from paragraph 7 to paragraph 12 analytically. The following table displays the total scores given by NS and NNS raters for six identical paragraphs.

Table 15

Statistical Analysis of the Total Scores the Analytic Assessment

	$\bar{x}$	SS	Median	Min	Max	U	p
NS	180,17	13,26	179,00	163,00	201,00	1,121	0,310
NNS	169,00	20,10	161,50	148,00	196,00		

It can be seen in the table that the standard deviations are 13.26 and 20.10 respectively. Although NS scored slightly more consistently than NNS, with the smaller standard deviation level, these deviations did not represent any significant difference in rater consistency between NS and NNS scorings. Regarding the minimum and maximum scores, it appears that NNS raters were slightly more severe, while NS raters were more lenient in analytic assessment. The lowest total minimum and maximum scores (148 and 196) belonged to NNS, while the higher ones were of NS (163 and 201). To analyze how the raters perceive the criteria and to identify any meaningful variance in analytic scoring, the six paragraphs in the second part will be demonstrated in the following section. The assessment criteria depended on certain categories were defined by the analytic rubric which were content, organization, use of language and mechanics. Content represented if the task was achieved and how well the ideas were developed or exemplifies. Organization standed for coherence, logical flow of the ideas and fluent sentence linking. Use of language meant grammatical and syntactical structures. Finally, mechanics was for spelling, capitalization and punctuation.

Table 16

Analytic Scoring of the Paragraph 7

Paragraph 7		$\bar{x}$	SS	Median	Min	Max	Z	p
Content	NNS	3,40	,84	3,00	2,00	5,00	-0,360	0,719
	NS	4,20	,79	4.00	3,00	5,00		
Organization	NNS	2,90	,99	3,00	1,00	4,00	-0,258	0,769
	NS	3,10	1,10	3,00	1,00	5,00		
Grammar	NNS	2,60	,52	3,00	2,00	3,00	-0,571	0,568
	NS	2,70	,67	3,00	2,00	4,00		
Vocabulary	NNS	3,30	,95	3,00	2,00	5,00	-1,149	0,250
	NS	3,10	,88	3,00	2,00	5,00		
Mechanics	NNS	2,80	,92	3,00	2,00	5,00	-1,148	0,251
	NS	3,20	,92	3,00	2,00	5,00		

Mann-Whitney U Test, $p<0,05$

Table 15 shows whether the assessment of the content, organization, grammar, vocabulary and mechanics categories of Paragraph 15 showed a statistically significant difference between NNS and NS, and it was found that there was no significant difference ($p>0.05$). NNS and NS made similar evaluations in all categories of Paragraph 15. When

the table is examined, it is seen that in the categories of content, organization, grammar and mechanics, NS made a slightly lenient assessment than NNS, but the difference between the two groups is not significant.

Table 17

Analytic Scoring of the Paragraph 8

Paragraph 8		x̄	SS	Median	Min	Max	Z	p
Content	NNS	3,10	,99	3,00	2,00	5,00	-1,148	0,251
	NS	3,60	,97	3,50	2,00	5,00		
Organization	NNS	2,90	1,37	3,00	,00	5,00	-1,301	0,193
	NS	3,70	1,06	3,50	2,00	5,00		
Grammar	NNS	2,70	,82	2,50	2,00	4,00	-0,606	0,545
	NS	3,00	1,05	3,00	2,00	5,00		
Vocabulary	NNS	2,60	,70	2,50	2,00	4,00	-1,895	0,058
	NS	3,20	,63	3,00	2,00	4,00		
Mechanics	NNS	3,00	,82	3,00	2,00	4,00	-1,545	0,165
	NS	3,60	,70	3,50	3,00	5,00		

Mann-Whitney U Test, p<0,05

Table 17 shows whether the assessments of the content, organization, grammar, vocabulary and mechanics categories of Paragraph 8 showed statistically significant difference between NNS and NS, and it was found that there was no significant difference (p>0.05). NNS and NS made similar assessments in all categories of Paragraph 8. When the table is examined, it is seen that in the categories of content, organization, grammar, vocabulary and mechanics, NS assessed slightly more leniently than NNS, but the difference between the two groups is not significant. In the organization category for Paragraph 8, the minimum evaluation score of NNS was 0, while the minimum evaluation score of NS was 2. This result shows that NNS made a harsher assessment in the organization category for Paragraph 8 (Table 16).

Table 18

Analytic Scoring of the Paragraph 9

Paragraph 9		$\bar{x}$	SS	Median	Min	Max	Z	p
Content	NNS	4,10	,74	4,00	3,00	5,00	0,000	1,000
	NS	4,10	,74	4,00	3,00	5,00		
Organization	NNS	4,00	,47	4,00	3,00	5,00	-1,734	0,083
	NS	3,50	,71	4,00	2,00	4,00		
Use of Language	NNS	3,80	,42	4,00	3,00	4,00	-2,737	0,006*
	NS	2,90	,74	3,00	2,00	4,00		
Vocabulary	NNS	3,50	,97	3,00	2,00	5,00	-1,290	0,197
	NS	3,00	,47	3,00	2,00	4,00		
Mechanics	NNS	4,20	,92	4,50	3,00	5,00	-1,540	0,123
	NS	3,50	,97	4,00	2,00	5,00		

Mann-Whitney U Test, p<0,05

Table17 shows whether the assessments of the content, organization, grammar, vocabulary and mechanics categories of Paragraph 9 showed statistically significant differences between NNS and NS, and a significant difference was found only in the grammar category ($p<0.05$). NNS and NS made similar assessments in the content, organization, vocabulary and mechanics categories of Paragraph 9. In the content category, both groups made the same assessment and gave a minimum score of 3 and a maximum score of 5. In the organization, vocabulary and mechanics categories, NNS made softer assessment than NS. While NNS gave a minimum score of 3 in the organization, vocabulary and mechanics categories, NS gave a minimum score of 2. There was a statistically significant difference between the scores of NNS and NS in the grammar category ($p<0.05$). The mean grammar score of NNS was 3.80±0.42, while that of NS was 2.90±0.74. NNS gave higher scores. When the standard deviation values were analyzed, it was seen that the standard deviation value of the NNS was lower. This shows that the assessments made by NNS are closer to each other. NNS gave scores of 3 and 4, while NS of 2,3 and 4.

Table 19

Analytic Scoring of the Paragraph 10

Paragraph 10		$\bar{x}$	SS	Median	Min	Max	Z	p
Content	NNS	2,80	1,32	3,00	1,00	5,00	-2,213	0,027*
	NS	4,10	,99	4,00	2,00	5,00		
Organization	NNS	2,60	,97	3,00	1,00	4,00	-1,204	0,229
	NS	3,30	1,06	3,00	2,00	5,00		
Use of Language	NNS	3,50	,53	3,50	3,00	4,00	-0,213	0,831
	NS	3,40	,70	3,50	2,00	4,00		
Vocabulary	NNS	3,00	1,05	3,00	1,00	4,00	-0,834	0,405
	NS	3,50	,97	3,00	2,00	5,00		
Mechanics	NNS	3,60	,84	4,00	2,00	5,00	-1,088	0,277
	NS	4,00	,67	4,00	3,00	5,00		

Mann-Whitney U Test, p<0,05

Table 18 shows whether the assessments of the content, organization, grammar, vocabulary and mechanics categories of Paragraph 10 showed statistically significant differences between NNS and NS, and a significant difference was found only in the content category ($p<0.05$). NNS and NS made similar assessments in the organization, grammar, vocabulary and mechanics categories of Paragraph 10. In the organization, vocabulary and mechanics categories, NS assessed more leniently than NNS. In the content category, there was a statistically significant difference between the scores of NNS and NS ($p<0.05$). The mean content score of NS was 2.80±1.32, while that of NNS was 4.10±0.99. NS gave higher scores. When the standard deviation values were analyzed, it was seen that the standard deviation value of NNS was higher. This shows that the assessments made by 20 NNS are more distant from each other. NS were closer to each other. While NNS gave a minimum score of 1, native English speakers gave a minimum score of 2.

Table 20

Analytic Scoring of the Paragraph 11

Paragraph 11		$\bar{x}$	SS	Median	Min	Max	Z	p
Content	NNS	4,20	,63	4,00	3,00	5,00	-0,336	0,737
	NS	4,00	,94	4,00	2,00	5,00		
Organization	NNS	3,90	,57	4,00	3,00	5,00	-0,583	0,560
	NS	3,60	1,07	4,00	2,00	5,00		
Use of Language	NNS	4,10	,57	4,00	3,00	5,00	-2,171	0,030*
	NS	3,50	,53	3,50	3,00	4,00		
Vocabulary	NNS	4,10	,57	4,00	3,00	5,00	-1,594	0,111
	NS	3,70	,48	4,00	3,00	4,00		
Mechanics	NNS	3,70	,95	4,00	2,00	5,00	-0,502	0,616
	NS	3,90	,57	4,00	3,00	5,00		

Mann-Whitney U Test, p<0,05

Table 19 shows whether the assessments of the content, organization, grammar, vocabulary and mechanics categories of Paragraph 11 showed a statistically significant difference between NNS and NS, and a significant difference was found only in the grammar category ($p<0.05$). NNS and NS made similar assessments in the content, organization and vocabulary categories of Paragraph 5. In the content, organization and vocabulary categories, NNS assessed slightly more leniently than NS. There was a statistically significant difference between the scores of NNS and NS in the grammar category ($p<0.05$). The mean grammar score of NNS was 4.10±0.57, while that of NS was 3.50±0.53. NNS gave higher scores. However, when the standard deviation values were analyzed, it was seen that the standard deviation value of the NNS was lower. This shows that the assessments made by 20 NS is closer to each other. NNS gave a minimum score of 3 and a maximum score of 5, while NS gave scores of 3 and 4.

Table 21

Analytic Scoring of the Paragraph 12

Paragraph 12		$\bar{x}$	SS	Median	Min	Max	Z	p
Content	NNS	3,60	,84	4,00	2,00	5,00	-2,062	0,039*
	NS	4,30	,48	4,00	4,00	5,00		
Organization	NNS	3,30	1,16	3,50	1,00	5,00	-1,352	0,176
	NS	4,00	,82	4,00	3,00	5,00		
Use of Language	NNS	2,80	,79	3,00	2,00	4,00	-1,812	0,070
	NS	3,40	,52	3,00	3,00	4,00		
Vocabulary	NNS	2,80	,63	3,00	2,00	4,00	-2,317	0,021*
	NS	3,50	,53	3,50	3,00	4,00		
Mechanics	NNS	3,40	,52	3,00	3,00	4,00	-0,872	0,383
	NS	3,60	,52	4,00	3,00	4,00		

Mann-Whitney U Test, p<0,05

Table 20 shows whether the assessments of the content, organization, grammar, vocabulary and mechanics categories of Paragraph 12 showed a statistically significant difference between NNS and NS, and a significant difference was found in the content and vocabulary categories ($p<0.05$). NNS and NS made similar assessments in the organization, grammar and mechanics categories of Paragraph 12. In the organization, grammar and mechanics categories, NS assessed more leniently than NNS. There was a statistically significant difference between the scores of NNS and NS in the content category ($p<0.05$). The mean content evaluation score of NNS was 3.60±0.84, while that of NS was 4.30±0.48. NS gave higher scores. When the standard deviation values were analyzed, it was seen that the standard deviation value of NS was higher. This shows that the evaluations made by 20 NNS are more distant from each other. NS were closer to each other. While NNS gave a minimum score of 2, NS gave a minimum score of 4. In the vocabulary category, a statistically significant difference was found between the scores of NNS and NS ($p<0.05$). The mean vocabulary score of NNS was 2.80±0.63, while that of NS 3.50±0.53. NS gave higher scores. When the standard deviation values were analyzed, it was seen that the standard deviation value of the NNS was higher. This shows that the evaluations made by 20 NNS are more distant from each other. NS were closer to each other. While NNS gave a minimum score of 2, NS gave a minimum score of 3. To summarize, when all the 6 analytically graded paragraphs were considered, it can be concluded that there is no significant difference between the NS and NNSs' scorings in

terms of rater consistency and scoring behavior in majority of the criteria. In paragraphs 7 and 8, no significant difference detected. There were significant differences in grammar criteria of the paragraphs 9 and 11. NNS were more lenient in their scorings and NS were severe for these two paragraphs. The content criteria of the paragraphs 10 and 12 were scored significantly different by the NS and NNS. In both of them NS assessed more leniently and NNS were more severely. In addition, there was another significant difference in vocabulary criteria of the paragraph 12. NS were more lenient and NNS were severe in their scoring behavior.

Thus, in the scoring analysis of all the 12 writings, It is understood that NS and NNS generally scored these writings in the manner in rater consistency and severity. The majority of the differences were not considerable differences. In the first part of the analysis, which consists of the first six paragraphs (1-6) scored through impression method. Only the paragraph 6 scored more severely than NS. In the second part which includes last 6 paragraphs (7-10) that were scored through analytic method, with the help of a rubric. Content and grammar criteria were scored significantly differently for two times. For content, NNS were more severe while NS were more severe for grammar. Additionally, vocabulary criteria was scored differently for once and NNS were more severe in their scoring behavior. As for consistency, in the impression scoring part NNS were slightly more consistent than the NS but there was no significant difference. Similarly there were not any significant difference in the analytic scoring.

According to the statistics, the NS and NNS had quite slight differences between each other in general. Nonetheless, the statements collected from NS and NNS raters during think-aloud sessions conveyed some expressions that were either severe or lenient. The think-aloud sessions and interviews demonstrated that the raters' perspectives differed in terms of severity. NS were tolerant against the students' errors in writings. The students' mistakes varied, despite the descriptive results showing that both raters assessed the students equally. In addition, NS adopted a more lenient approach toward the mistakes, praising the students' attempts to employ frameworks that are more intricate and taking risks. It is not noticeable from the scorings but pursuing errors in a text may result in misjudgment and may cause us to develop bias against their performance. Some rater's approach is quite nice and lenient as it is obvious from the following quotes:

> *"The content is fairly good, it's difficult not to be kind of sidetracked with the the errors that are there, so obviously you know a challenge as a rater is you know you're looking at the errors and sometimes the errors can mislead you. So looking*

at the errors. The quality is not really appropriate, but for a low level, it's quite nice that they're trying to use a different word the student is trying to use a wide range of ideas and that's why she's making lots of mistakes." (NS 7)

"There are mistakes. There is a number of mistakes, but then there is also some impressive use. You know, you ever have that problem with grading where sometimes they try harder so they make more mistakes, but you are proud that they tried. I feel like it is one of those. They made more mistakes, but they also tried some things that most students wouldn't try." (NS 10)

"Their languages they've tried to use more complex language, and they've made more mistakes. So overall, I would say that this essay demonstrates. A better command of English, even though it's probably got more inaccuracies in it." (NS 1)

However, NNS had a severe approach to the students taking risks and their attempt to use complex sentences that they had not mastered and could not use properly. They supported the idea that they should try not to include anything wrong or informal in their writing and they were even less tolerant when the students made basic mistakes that are under their level:

"He wants to say 'I think walls should be soundproof', but if he can't use the structure correctly, he shouldn't write it at all." (NNS 9)

"While finishing, he made up something called 'discripting' and his spelling is incorrect. Okay, this might be a bit of a difficult word, but he could have just not written it at all. I think he's pushing his luck with this sentence, which also bothers me." (NNS 1)

"When writing 'responsibility', they made a lot of spelling mistakes. I find this very bothersome as well. For example, if they cannot spell a relatively easy word correctly, it makes me wonder. It feels like this student does not really know what they are doing. Because they cannot even spell this word correctly, I think. Then, this could also affect the overall writing score I give. Especially for a simple word. If they spelled it wrong, it directly affects my judgment." (NNS 5)

In parallel with these findings, when the transcribed data analyzed, it was discovered that the NNS raters' language is more critical and they used harsher expressions.

"It annoys me when I see the mistake in the mistake in the sentence. I mean, he shouldn't say 'wheathe'r instead of weather. Is this a high level student?" (NNS 1)

> *“I get irritated when they misused the word of ‘situation’. They do not know ‘the words of issue’ or ‘case’. This is why they do this mistake but it’s an important mistake for me because it hinders the meaning.1” (NNS 8)*

In contrast, the NS raters’ expressions are quite more understanding and adopted a sympathetic point of view as it is clear in their statements:

> *“But they haven't learned how to say it yet, so they're trying to say something which is over their level. I am learning French and this is me in French. I am saying stuff in a really basic way. But the idea is the idea is complex, but my language is basic. This is one of the biggest problems as well for the students. And we don't want to tell them stop being complex.” (NS 2)*

> *“Only really small errors here almost. Does not change the meaning, doesn't interfere with the meaning, just little mistakes. Probably that they keep making. Like if I show them this essay and if I say if I say: “What's the problem here? Probably they're going to go say “Ohh, It must be listen ‘to’ others. Sorry sir. OK.” (NS 4)*

> ***“****I'm not sure. But this I'm going to say a higher score. You know the mistakes that they made are the same mistakes that everybody makes usually. That's why we're students. They are learning how to do it.” (NS 6)*

In addition to student mistakes in the paragraph, it was reported in the interviews that misunderstandings of the writing task was also harshly punished:

> *“When the person has good English skills, but they misunderstood the prompts or you know something they just mixed up something by accident and the writing, the writing that they put out is really good, but we have to grade them down because of their misunderstanding. They misread the instructions or something like that. . Something less than like a 60%. I think that's just really discouraging.” (NS 4)*

All in all, the scorings of the NS and NNS were statistically analyzed and the findings were presented in this part. As the standard deviations considered, no significant difference between NS and NNS ratings was found in terms of rater consistency. There were only slight changes between the NS and NNS. It can also be concluded from the minimum and maximum scorings that the lower the quality of the paragraph, the lower the score and the more differentiated the scoring. In this case, as the paragraph quality decreases, rater consistency also decreases. The better the paragraph is written, the closer the scores are and the higher the rater consistency.

As for scoring behavior, there was no significant difference between NS and NNS raters except for a few. In the impression scoring part, only one paragraph significantly different and NNS raters were more lenient in their scoring. In the analytic scoring part, the rater's approach to assessment criteria was investigated. In most of the paragraphs, NS and NNS rated in parallel with each other apart from a few significantly different criteria. In those exceptions, NNS rated use of language more leniently than the NS for two times and NS rated content more leniently than the NNS for two times and vocabulary for once. Therefore, NS and NNS scoring behavior did not differentiate meaningfully in terms of severity. However, the statements from the think-aloud sessions and interviews implied that NNS raters' approach was more severe and critical to the students' errors whereas NS adopted an understanding and tolerant attitude.

Research Question 2: How do Native and non-native raters consider various criteria when evaluating writings? The primary objective of the second research question is to investigate the assessors' approaches to writing assessment and to distinguish between the writing assessment strategies of NS and NNS. Think-aloud, sessions and interviews were analyzed to uncover the rationale behind their assessment strategies. Through the identification of recurring themes, we aimed to observe the perspectives of both NS and NNS raters, with the intention of understanding the underlying logic guiding their writing assessment logic.

The think-aloud sessions were conducted to track the NS and NNS raters assessment strategy. They were asked to assess the 12 paragraphs first through impression scoring and then through analytic scoring articulating their thoughts and interpretations concurrently. It was intended to identify to observe their approach and some considerable differentiations were detected as well as common points on which most members of the two group agree.

Firstly, differences between the decision-making behaviors were investigated. For this, the raters' assessment strategies were classified according to 35 decision-making behaviors which were developed by Cumming and Sakyi (2002). It was revealed that both NS and NNS raters as shown in the table frequently used 23 of them. Hence, in this study the decision making behaviors were categorized under these 23 behaviors.

Table 22

23 Decision-Making Behaviors in Writing Assessment (adapted from Cumming & Sakyi, 2002)

Self-monitoring Focus	Rhetorical and Ideational Focus	Language Focus
Interpretation Strategies ● Read essay prompt ● Read/reread composition ● Scan composition *Judgement Strategies* ● Consider own personal response ● Compare with other writings ● Summarize ideas ● Articulate general impression	● Assess reasoning, logic, topic ● Assess relevance ● Assess coherence ● Assess creativity ● Assess text organization ● Assess style, genre	● Classify errors ● Assess comprehensibility ● Assess fluency ● Assess lexis ● Consider error frequency ● Consider gravity of errors ● Consider syntax and morphology ● Consider spelling and punctuation ● Edit phrases for interpretation ● Rate language overall

The 23 decision making behaviors that were presented in the table 21 were used by the NS and NNS raters in this research. They were divided into three subcategories focusing on self-monitoring of interpretation and judgement strategies, rhetorical and ideational strategies and language strategies. Self-monitoring focus strategies were about the way the rater interpret and judge the paragraphs. To start with the interpretation strategies, which referred to the first step of analyzing a paragraph, the both the NS and NNS raters read the essay prompt, read the composition itself and scan the composition . While scoring the paragraphs. In the second step, when they were judging the paragraph they considered and articulated what their response would be, compared the paragraph with the counterparts, summarized the ideas and articulated their general impression about the paragraph.

As shown in the table, the NS and NNS raters judged the paragraphs focusing on six strategies which are the rhetorical/ideational qualities of the paragraphs. The strategies and what they referrer to be explained as follows:

- *Assessing reasoning, logic and topic:* assessing the content development.
- *Assessing relevance:* considering the task achievement and the connection of the ideas in the text.

- *Assessing coherence:* considering the flow of the ideas, judging how well the ideas blended.
- *Assessing creativity:* considering the quality of the examples given and how strongly the arguments were supported.
- *Assessing organization:* considering the topic, supporting and concluding parts of the text.
- *Assess style, genre:* interpreting the manners of the lay of ideas, figurative language and form

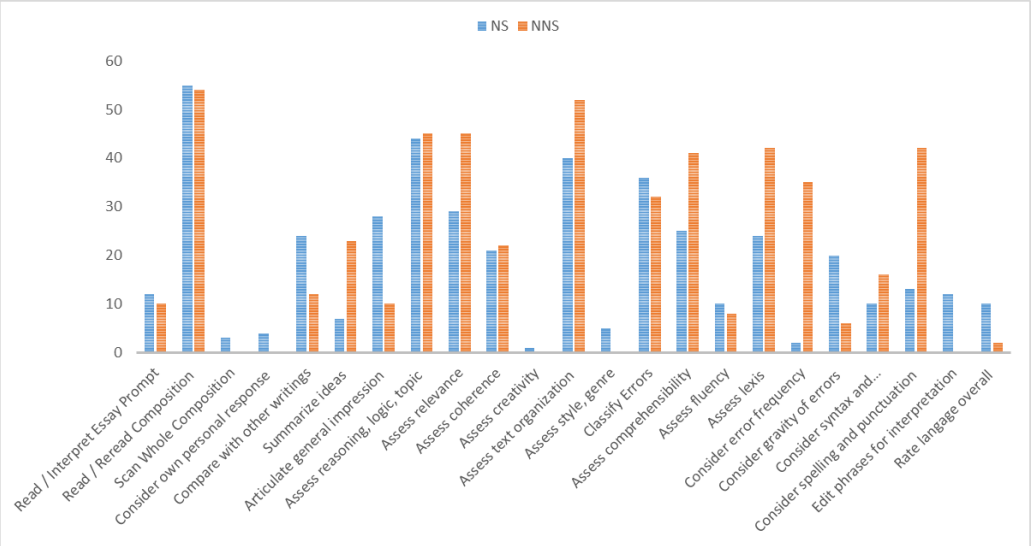
As for the language strategies, NS and NNS raters made use of 10 strategies. How they function is described below:

- *Classify errors:* categorizing language error types
- *Assess comprehensibility:* evaluating the meaning is intervened by the language mistakes
- *Assess fluency:* if language hinders the easy reading flow
- *Assess lexis:* evaluating the quality and variety of vocabulary
- *Assess syntax and morphology:* evaluating sentence structure and word form/structure
- *Consider error frequency:* detecting how often the errors occurs
- *Consider gravity of errors:* detecting how serious are the errors
- *Consider spelling and punctuation:* determining mechanics mistakes
- *Edit phrases for interpretation:* clarifying ambiguity of the phrases
- *Rate language overall:* giving a general impression about the language accuracy

In the table above, the frequency of decision-making behaviors of NS and NNS raters were displayed. The decision-making behaviors were coded on the think-aloud session transcripts and counted by the researcher. For the coding, the first 6 paragraphs (paragraph 1 - Paragraph 6 of impression scoring) were used on the grounds that a free scoring method would be more useful to identify the raters' personal approaches rather that analytic scoring method which directs the raters to evaluate certain criteria. In table 22, it is clearly seen how many times the members of the groups made use of a particular strategy. In addition the total numbers provided clues about the most important criteria for the raters or mostly-remarked points. Besides, the criteria that they overlook can also be distinguished.

Table 23

Frequency of Decision-Making Behaviors



The table shows that NS and NNS raters exhibit parallel behaviors in general. In other words, they make use of same assessment strategies almost equally except for only two, which are ‘considering error frequency’ and ‘considering spelling and punctuation’. According to the table, NS raters who considered error frequency for only a couple of times whereas NNS employed this strategy more than thirty times. In the interview part, it was clearly stated that the raters perceive the assessment more fair and accurate if they count the mistakes as quoted below:

> “I do care that the student repeats the same mistake many times. Sometimes I assess similar mistakes in one group and sometimes I count them one by one.” (NNS 3)

Similarly, spelling and punctuation was considered far more times than NS. These numbers are 42 and 14 times for NNS and NS respectively. Based on this information, It can surely be concluded that NS and NNS exhibited the same decision-making behaviors in the same density apart from ‘considering error frequency’ and ‘considering spelling and punctuation’. These two were performed by NNS, far more than NS.

The raters’ statements quoted from think-aloud sessions supported these findings. NS raters put little emphasis on mechanics with the expressions such as ‘There is just a typo’ or ‘It’s a basic spelling mistake but that’s Ok. I do that too.’ On the contrary, NNS raters expressed that they perceive mechanics important and they consider the mistakes in mechanics which refers to spelling and punctuation mistakes due to the reason that

students never care about these rules even though they are important elements of a written text:

> '*When there are problems such as not using punctuation marks properly and not spelling words properly, I deliberately lower the score because I want to convince the students that this is important in language. So I cut off points especially here so that it is a deterrent and they have to learn it.' (NNS 1)*
>
> '*I deducted a point again because there were no mechanical commas or anything. If we don't, they don't care about it as a part of written language. Mechanics actually do not deserve so many points, but it creates a big problem. Some of them may causes a change in meaning in the grammatical structure*.' *(NNS 6)*

The raters were asked to choose the least and the most important factors in their opinion to be able to identify if their statements meet their practice the think-aloud sessions. However, the NNS raters' expressions contradicted with their practice demonstrated above. In the interview part, the raters were asked a question about the least important criteria for a good writing and 85 % of the raters (both NS and NNS) choose mechanics. To restate, there was a significant agreement between the groups on the idea that mechanics is the least important criteria as depicted in the following figure 8:

Figure 8

The Least Important Criteria

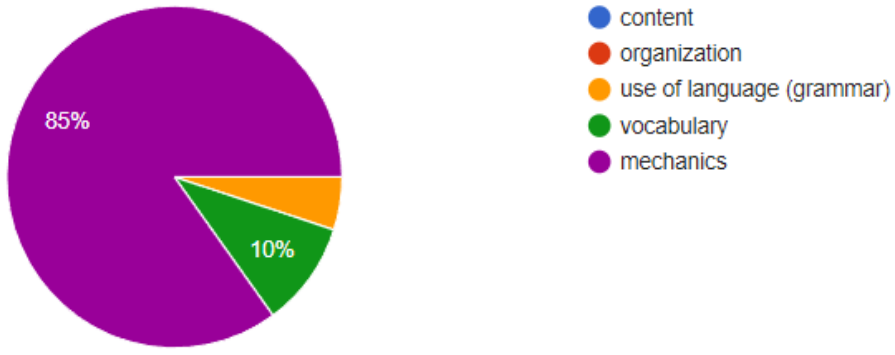
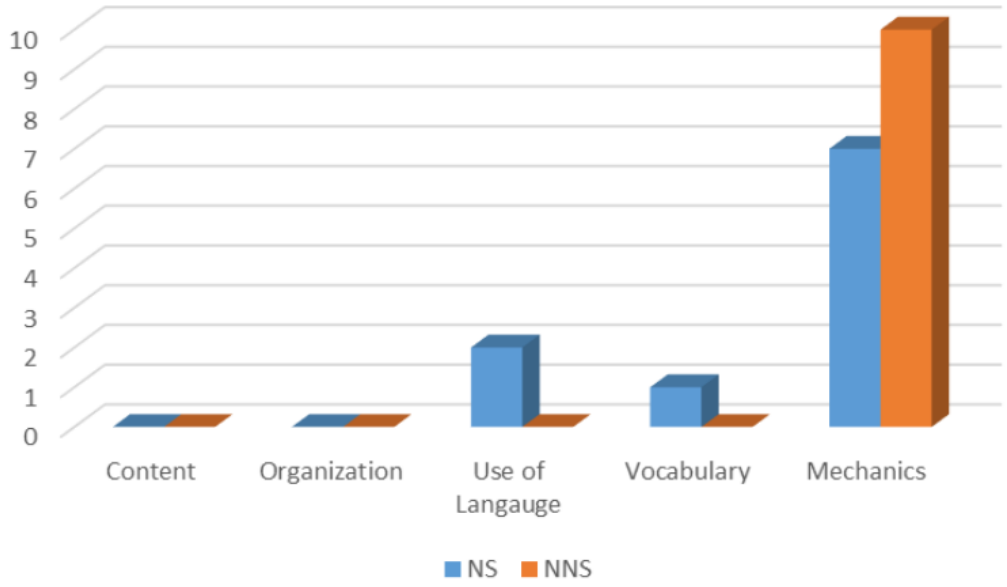


Table 24 below presents detailed information on the quantity of votes for the least important criteria. According to the table, the majority of NS raters perceived mechanics as the least important criterion. Specifically, 7 NS labeled mechanics as the least important, while 2 of them chose the use of language and 1 selected vocabulary. Interestingly, all 10 NNS agreed that mechanics is the least important criterion and did not choose any other criteria. Table 22 also suggests that spelling and punctuation mistakes (which refers to

mechanics) were considered by both NS and NNS raters. However, NNS assessed these mistakes 42 times whereas NS did it for 13 times. Thus, their practice and their statements about the importance of mechanics did not overlap. These results are contradictory to both their statements and their decision-making behavior during practice sessions. Conversely, NS practice during think-aloud sessions and their statements in interviews show more alignment compared to the NNS.

Table 24

The Least Important Criteria for the Raters



In addition to the least important criteria for themselves, the raters were also asked the most important criteria in the interview questions as well. Their choices are depicted in the pie chart in the figure 9 below. 50% of all raters chose content, %35 of them chose use of language and %15 of them chose organization. In that majority of the raters thought that, content and organization are the most important criteria for a high-quality paragraph.

Figure 9

The most important criteria for raters

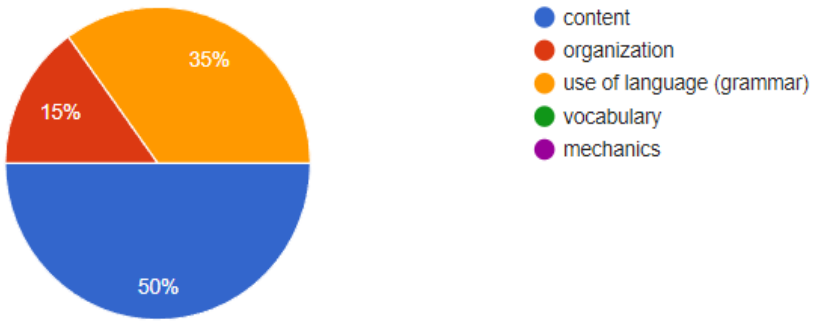
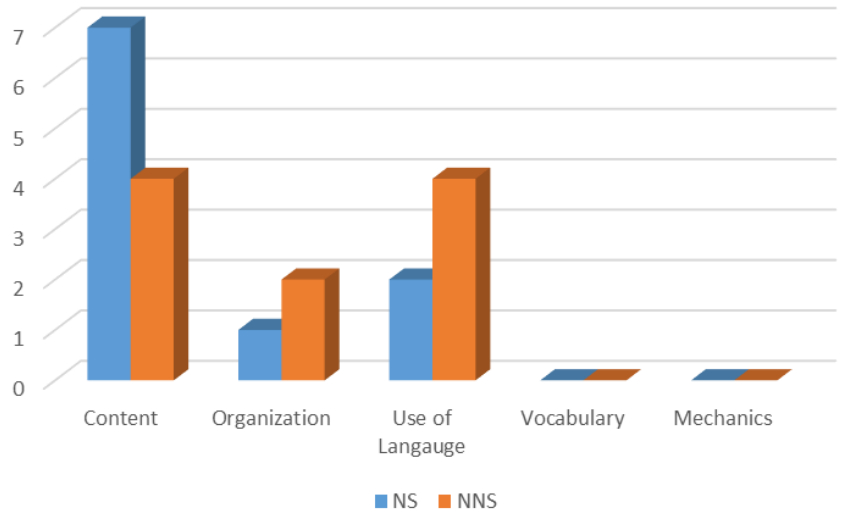


Table 23 shows which criteria is the most important for the NS and NNS raters. Most of the NS selected content while NNS made different choices. 7 NS voted for content, 2 for use of language and 1 for organization. As for NNS, content and use of language were equally picked by 4 raters, and 2 of them went for organization. Considering this table, it is clearly seen that the least number of NNS raters (only 2) preferred organization as the most important criteria. However, in table 22 above, we can also see that assessing text organization was the mostly used strategy by NNS, even more times than content. Here, there is another contradiction between NNS's claims and their practice.

Table 25

The Most Important Criteria for the Raters



The statements provided by NNS raters during interviews supported their emphasis on organization in evaluating texts. They explained why they prioritize organization when scoring writings as follows:

> *"This is a good paragraph because he has written the organization as it should be. I think like this, because if the student understands how to do the organization, the content is already automatically enriched. Because after the supporting idea and if he thinks that I need to write examples, he already opens that topic directly or he is not confused. So it goes to a good place. In order to write well, he inevitably needs to know the organization." (NNS 1)*

> *"I pay more attention to organizational things in writing. Does the child have the ability to build organization in his/her head? This is more important." (NNS 5)*

> *"I think organization is one of the most important things. That comes before grammar for me. Because grammatical mistakes always happen. Every student makes them, but if his organization is good, he actually knows what a paragraph consists of and I think he can write good things." (NNS 8)*

The NS and NNS raters expressed their ideas about the most important criteria in writing assessment and their assumptions about which group prioritize which criteria in the think-aloud sessions. One of the most remarkable assumptions remarked by the participants was that NS raters would care more about content while NNS would consider grammar. It can be concluded looking at the tables illustrating decision-making behaviors and the most important criteria. As the decision-making strategies suggested, NNS paid attention to grammar accuracy. In the think-aloud sessions it was revealed that NNS referred to basic grammar errors such as to be (am, is, are) mistakes much more frequently than the NS did. However, when the grammar mistakes started to intervene the meaning both of the groups were reacted to the errors almost equally:

> "*While there are a few grammatical errors in this paragraph, they are not significant enough to detract from the overall meaning." (NNS 1)*

> *"I found this paragraph quite challenging to understand. While I'm attempting to grasp the content, the fragmented sentences are hindering my comprehension. They disrupt the flow of meaning and make it difficult to discern the intended message." (NNS 10)*

> *"Even if he has grammatical mistakes or misspells words, the important thing is to get the message across; I never understood this paragraph because he used grammar incorrectly to the point of distorting the meaning. The content suffers because the writer cannot get the message across." (NS 2).*

According to the statistics above, NNS raters are obviously more sensitive and severe against basic grammar errors; nevertheless, this did not surpass the weight of meaning in the text. In other words, NNS did not ignore the content and meaning nor put the grammar before them, neither did the NS raters. Both NS and NNS prioritize general comprehensibility and clarity in a paragraph when scoring it. They expressed their attitude in the quoted below:

> *"This paragraph is weak because it burdens the reader with excessive effort to understand. The reader must spend significant mental energy to decipher its meaning." (NNS 2)*

> *"Not completely, but yeah, we have to work quite hard to understand what the person meant, It's a little bit hard to follow then. Content is great, but we have to work really hard to understand this paragraph, so I'm sorry it needs lots of work." (NS 6)*

> *"I think it is really important to consider how much work the reader has to do to make sense of what they're saying, what they wrote." (NS 9)*

In terms of decision-making behaviors, the raters' assumptions were partially true. NNS placed greater emphasis on basic grammar errors and their frequency, while NS prioritized the gravity of errors. NS also showed slightly more concern for the content and meaning of the paragraphs. However, both NS and NNS held similar attitudes towards language and grammar errors, especially when they intervene in the meaning.

In addition to think-aloud sessions, the raters were posed 5 interview questions. These questions were as follows in order:

1. *Do you enjoy assessing writing?*
2. *Do you encounter any challenges or difficulties when assessing writing?*
3. *Which scoring method is more useful in your opinion? Why?*
4. *Do you think writing are assessed fairly today?*
5. *Do you think it is more reliable when the writings are assessed by native raters?*

In response to the first research question, raters were queried about their enjoyment of the writing assessment process. A majority of raters expressed a lack of enjoyment, citing

stress associated with maintaining fairness in their evaluations. Consequently, raters reported finding the task of providing feedback on student writing to be less difficult. Assessing writing is viewed as a reflection of students' abilities, which presents its own set of challenges. Furthermore, many raters noted that aligning the writing with the rubric, a critical step for ensuring fairness in assessment, is particularly daunting due to its time-consuming nature as they mentioned in the interviews:

> *"When I merely give feedback to the writings, I can talk about it with them and help them develop their ideas, and then we can look at the language that they've used. But when you're assessing, it's strategically different because then everything you have to take all the characteristics of the writing and match them to the different criteria in the rubric." (NS 2)*

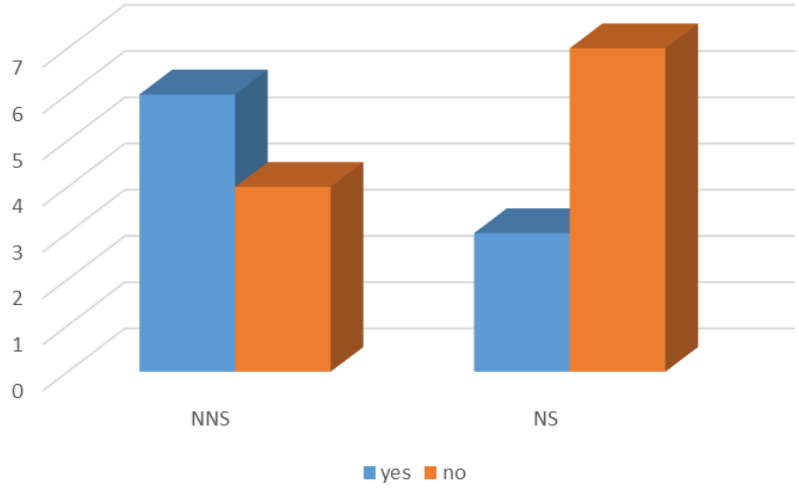
> *"I think it is quite stressful because you'd be spending a whole day marking writing papers and of course you'd get tired." (NS 7)*

> *"It is a stressful process for me. If I give homework in class and the students are doing it for their own improvement, I give feedback and if we are just doing it for practice, I enjoy it. I like giving feedback but if I am doing it for an exam, I definitely get nervous because I have a hard time trying to be fair. (NNS 1)*

In total, 9 NS and 9 NNS declared that they enjoy writing assessment for they perceive it stressful and tough and the reason is that it is too challenging to assess them fairly and its tiring to match the rubric with the writings while only one NS and one NNS reported that they enjoy it. Moreover, depending on their answers regarding that they find it time-consuming, the raters were asked about their tendency to finish scoring the writings as soon as possible. Most of the raters stated that they try to finish scoring as quickly as they can. As it is seen in the table below, 9 raters said that they try to finish quickly while 11 of them said they spend time.

Table 26

The Raters' Tendency to Finish Writing Assessment



As indicated in this table, majority of the NNS tend to finish assessment as quickly as possible while most of the NS do not. When the total of the numbers considered regardless of their native language, 9 raters tend to complete assessments swiftly, whereas 11 raters expressed the opposite tendency.

The second research question was about the difficulties that the raters have during the writing assessment. All of both NS and NNS raters complained about the subjective and time-wasting nature of writing assessment. In the interviews, it is clearly conveyed that they have the dilemma of assessing too generously or too harshly and concerned about being able to fair in their scorings and this led them to reread and rerate the paragraphs as explained below:

> *"Trying not to be subjective is difficult. I think you saw me hesitating like is this a four or a three when I was scoring the paragraphs. I don't know if it's just me being harsh or generous. That's the thing that we tried to get rid of and I'm not sure if we can ever completely get rid of it. That's the problem." (NS 6)*
>
> *"Even though I follow a rubric, I am never sure whether I have given the right grade or not. I occasionally want to go back and look at the papers from time to time because I wonder if I give someone a higher grade and someone a lower grade. That is a very time-wasting process for me." (NNS 1)*
>
> *Sometimes it bothers me that there might be a lot of discrepancy between those papers. I feel the need to go back and check if there is too much difference between the first paragraph you read and the last paragraph I read. Or, for example, I give a*

point to a student. Then I think if I could have been more tolerant to the previous student. I think about how I can minimize the discrepancies between the notes, that is, how I can read completely precise. Sometimes I look at it several times like this. So trying to be fair is very difficult. (NNS 2)

Both NS and NNS raters mostly claimed that using a clear rubric helps them to be more objective in their assessments. They suggested that if the rubric items were described plainly and the errors were explicitly categorized, they felt much more comfortable while scoring the writings:

"If the criteria like the rubric has kind of crossovers from the different sections and it isn't clear enough exactly how separate that is, then that's problematic. But if the rubric is really, really clear then I feel really confident. But if the rubric is telling me, where I have to mark, I don't want to penalize the student twice in two different sections. So, I think if you have a good one, then your job is quite easy." (NS 3)

"If the rubric is sufficiently directive and has very clear instructions, for example, if it says that you can't give more than one from content and no more than three from the others, I don't find it difficult to evaluate. I can go straight to the outcome." (NNS 2)

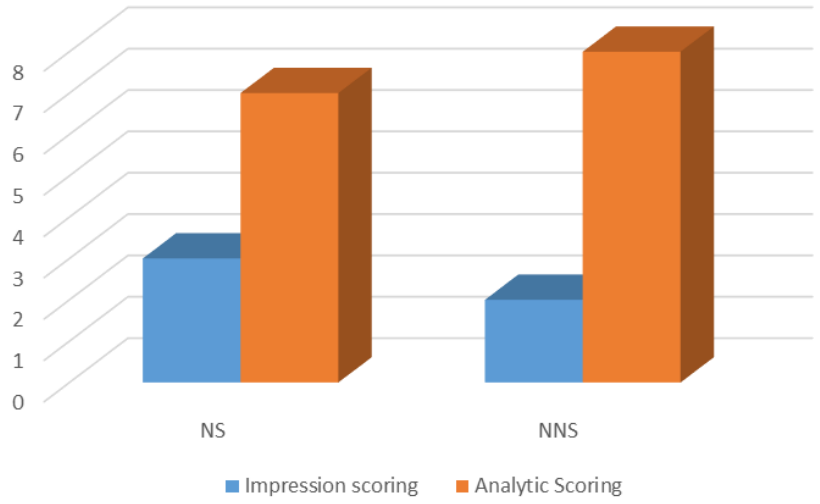
"Writing assessment is subjective. I need a very good rubric. It shouldn't be open to interpretation. For example, the rubric says that a few mistakes are acceptable. For example, they should explain what they mean by a few mistakes in ruby. I think I should be able to understand how many are too many and how few are too few when I read it." (NNS 4)

For example, the rubric says adequate and appropriate. I don't understand what it says is adequate and what it says is appropriate. When the explanation is not enough, I still evaluate according to my own opinion. Rubrics should be more descriptive, it should be described step by step what is wanted. (NNS 9)

To summarize, most of the NS and NNS raters share the same or very similar concerns about writing assessment. Most importantly, it was reported that they doubt they could score the writings fairly and they considered using an analytic rubric would be useful both to be fair and practical in terms if time and energy saving on the condition that the rubric would be a well-structured one.

Table 27

Scoring Method Preference of the NS and NNS raters



Depending on their thoughts on the rubrics above, the raters were asked about their preference of impression assessment or analytic assessment method. Not surprisingly, most of the raters asserted that analytic assessment using a rubric is more practical and more reliable. 7 NS and 8 NNS stated that they would prefer analytic assessment with a rubric while 3 NS and 2 NNS indicated that they find holistic method more effective as shown in the table above.

Considering the total assessments of the two groups, the implementations of the two methods diverted in scores. The table below compares minimum and maximum of total scores of both the impression and the analytic methods of scoring used by NS and NNS. It includes standard deviation value for both.

Table 28

The Total scores of Impression Assessment and Analytic Assessment

	Total Scores of Impression Assessment			Total Scores of Analytic Assessment		
	SS	Min	Max	SS	Min	Max
NS	43,19	105,00	209,00	13,26	163,00	201,00
NNS	39,14	116,00	220,00	20,10	148,00	196,00

Regarding impression assessment standard deviation of the NS ratings was 43,19 and NNS was 39,14. As for analytic assessment, the standard deviation of the NS 13,26 and NNS was 20,10. In that, the standard deviation of impression assessment was considerably higher than the analytic assessment. These statistics emphasize that the scores of impression method had a doze of analytic scoring and the reliability is lower in impression assessment. Both NS and NNS raters made similar interpretations about the reliability of this method:

> *"Analytic scoring do not let huge discrepancies. And how? The teachers do not decide what grades to give. And I think it was much better for the students and the teachers." (NNS 7)*

> *"When I score through impression method, It isn't systematically rated at all. I mean, these are just random numbers really." (NS 1)*

Apart from that, minimum and maximum scores of the impression assessment and the analytic assessment also diverted noticeably. The lowest total scores awarded by NS and NNS in impression assessment were 105 and 116 respectively whereas they are 163 and 148 for the analytic assessment. When it comes to maximum scores, NS gave 209 and NNS rated 116 the most in the impression scoring. In the analytic scoring, the NS group scored 201, while the NNS group scored 196 at the utmost level. To conclude, the total maximum scores did not remarkably vary whereas there was a substantial difference between the total minimum scores. The minimum scores of the analytic assessment were notably higher than the impression scoring. Some of the raters also expressed that either they had to give a higher mark unintentionally or revised the score depending on their impression. Thus, they combine both methods:

> *"Both methods have advantages and disadvantages. Impression methods help me focus on content, and analytic method put emphasis on the other criteria and hinder me from forgetting about one criteria." (NNS 4)*

> *"Maybe it is healthier if I add my holistic point of view and write a total score at the end while following rubric." (NNS 10)*

> *"When I score according to the rubric, for example, I should give 20, but you look, it's not good enough to get 20. I may have to give more than they deserve." (NS 7)*

> *"Sometimes, I change the rubric score increasing or decreasing, say, mechanics grade according to my holistic opinion to give them what they really deserve."*
>
> *(NS 3)*

The third interview question was very much about the raters' perceptions of the two methods of assessment, holistic (impression) and analytic. Analytic assessment included the guidance of a rubric. The raters carried out both of the methods and shared their ideas about these methods in the interview part. Firstly, the statistics of the total scores showed that analytic assessment was more reliable than the impression assessment. Most of the raters, NS and NNS, clearly stated that analytic assessment and the rubric guidance should be the primary method for it is more reliable. Furthermore, in addition to reliability some of the raters mentioned further advantages of the analytic method such as, decreasing discrepancies or drawing the attention to all categories and preventing the rater to forget scoring an important criteria.

> *"Without a rubric, things can be missed. For example, I may not pay much attention to vocabulary or grammar at that moment. I find myself looking at the content. I might miss mechanics or spelling because I read 20 papers on average." (NNS 3)*

Fairness in assessment was the main concern of the fourth interview question. The NS and NNS raters were asked if they doubted that they assess a paragraph fairly and are there any ways to develop today's writing assessment traditions. The main purpose of this question was to reach the insights of the NS and NNS opinions about the concept of fairness and benefit from their experiences for this study as well as detecting any differences between their points of views. First, majority of both NS and NNS raters accepted that writing assessment is a subjective process and it is not sensible to expect it to be fully objective:

> *"Writing assessment has to be subjective in a way." (NS 7)*

> *"We try to be objective but of course we cannot avoid subjectivity." (NS 5)*

However, they also came on terms that fairness in writing assessment can be further improved through some activities. Mostly, both NS and NNS referred to 3 main principles which are standardization, detailed rubrics and cross-marking. The raters who referred to these principles were demonstrated in the table:

Table 29

Suggestions from NS and NNS Raters on Developing Fairness

	NNS	NS
Standardization	NNS 1, NNS 2, NNS4, NNS 6, NNS 10	NNS 1, NNS 2, NNS 5, NNS 7,
Cross-marking	NNS 1, NNS 5,NNS 9, NNS 10,	NS 2, NS 3, NS 7, NS 8
Detailed-rubrics	NNS 1, NNS 4, NNS 6, NNS 8, NNS 9, NNS 10	NS 2, NS 4, NS 8, NS 9, SNS 10

In their statements, 'standardization' encompasses any activity or training aimed at establishing consistent criteria and behavior in large groups of raters during writing assessments. These activities, as mentioned by both NS and NNS of English, may include trainings, workshops, or conferences. Long-term standardization efforts, which include comprehensive training sessions, were suggested during interviews. 'Cross-marking' involves the assessment of writings by multiple raters to enhance reliability. NS and NNS raters indicated that they often collaborate with partners to cross-check each other's assessments for increased reliability, suggesting that anonymizing the names and scores of the writers could further enhance fairness. Additionally, many raters expressed dissatisfaction with the lack of detail and clarity in existing rubrics, emphasizing the need for their development to ensure fairness and reliability:

> *“It can minimize discrepancies, but still, the evaluation depends on what the scorer understands from the rubric. so it's subjective again. And that's why it has to be spot on.” (NNS 2)*

Overall, the findings suggest a convergence of views among NS and NNS raters on strategies for promoting fairness in writing assessment.

The last interview question targeted the differences between NS and NNS raters. The question was directly addressed to both NS and NNS participants, seeking their personal opinions developed from their experiences, as all of them had worked with the other group in their institutions.

Finally, the differences between NS and NNS raters’ assessment was investigated through the raters opinions and experiences in the interviews. The participants were asked if they had observed some general distinction between these two groups depending on their

professional life in terms of writing assessment strategies and behaviors. Initially, it was uttered that there is possibly no significant difference regarding reliability in assessment. All of both NS and NNS frankly stated that they did not believe any remarkable difference existed in the reliability of the assessment. Instead, all of the NS and NNS raters came to an agreement on teacher education and training in the most important factor in reliable, fair and accurate scoring in writing, not the native language:

> *"If their field is not English language education, I don't think the native speaker's evaluation will be reliable because they should have a training on what to look at, what to focus on. For example, if they have not taken an academic writing course, they may ignore important points. In other words, if we ask an ordinary person, they may not know the rules of academic writing, so they may not be able to assess correctly. So I don't think it's because of the difference in mother tongue, I think it's more about looking at the background of the person." (NNS 10)*

> *"I think it depends on the native speakers background and teaching experience. Like, if I never teach writing, then I might not be a good grader of a text." (NS 3)*

> *"I think you know it really depends whether you're a native or not a native. I don't think it matters, really. It really depends on whether you have been really trained well to be able to read students work." (NS 7).*

Apart from the education and background, the participants also addressed that necessarily mean having a thorough understanding of the language. In other words, someone could be very poorly educated in their own language, as the raters claim:

> *"Some of the questions that these native speakers ask about English, I mean they are so poorly educated about their own language. I'm just like, Oh my God. Really, you know, like, wow, it's just it's made me so cynical about the quality of education. I think, that's really a factor. I think that is probably is the most important factor. It's the level of education that the person who's assessing has. And I'm sure that if I shared that with any of the non-native colleagues who all learned English as a second language, they would all laugh and make jokes about how poorly, how poorly those natives speakers know their own language." (NS 1)*

Nevertheless, most of the participants also accepted that native language of the rater influence the assessment process. In that, they stated in the interviews that being a native English speaker or sharing the mother tongue of the writers have their advantages.

For instance, some NS and NNS indicated that NS raters are more knowledgeable in the authentic language use. Some NS stated that they help their NNS colleagues with some idioms or some phrases about the most natural ways to say it, when assessing writing:

> *"Sometimes my coworkers. If they're not sure about something, they ask me or someone else and they say it's OK, right? And I say, not really." (NS 6)*

> *"I may find a word appropriate for the context I want to use it in, but a native speaker may say that this word is not used in that sense." (NNS 4)*

In the think-aloud sessions NS and NNS some similar situations had occurred. NS and NNS commented of differently. The writers made mistakes depending on their native language while writing their paragraphs. This type of mistakes are called negative transfer from the native language, which is Turkish in this case. To illustrate, a statement from paragraph 9 can be analyzed. In this paragraph, the writer mentioned his dream hotel and to signify that he prefers quiet rooms he stated '*the walls shouldn't pass the voice*'. However, this statement contradicts with authentic use of English language. To put it another way, the writer intended to express an idea over his language competency and influenced by his native language while trying to write it. This statement was one of the examples that is negatively transferred. Most of the NS raters notice this statement and interpreted it whereas only 3 NNS recognized it:

> *"Some bad syntax or spelling errors or like the 'walls shouldn't be past the voice', so he's got the word voice in there and pass like we understand what she means, but it's grammatically inaccurate. (NS 1)*

> *"This is because they don't know which word to use. The walls. It needs to be kind of a longer clause of the sentence. English doesn't have a nice way to say this so. He could have said 'We shouldn't be able to hear people's voice through the wall' or 'the walls shouldn't be thin. Because if they are thin, we can hear other voices.' (NS 5)*

According to the interviews, both NS and NNS also stated that sharing the same native language with the writers or knowing their language to some extent, help the raters to identify the errors:

> *"We can empathize more because the students' mother tongue is the same as ours, Turkish. Some of their sentences are grammatically wrong, but we can understand them in English. Therefore, the sentence does not sound problematic for us. Natives*

may not understand it while reading and this may be reflected in the scores." (NNS 8).

"For the Turkish teachers, reading Turkish students, they know exactly what the student means because it makes sense to them, like from Turkish to English. A Turkish teacher would know what they're saying, but as a native speaker teacher, sometimes I wouldn't know." (NS 4).

"There might be a difference in terms of sharing the same mother tongue. But that's nothing to do with being a native speaker. I mean, if it was a Japanese person assessing the English or an English person assessing the English, I don't think it would matter. But if it was a Turkish teacher assessing Turkish learners, it is possible that they would recognize impressions and things and say; oh yeah, I know what they're meaning." (NS 1)

There were some cases in this study that prove the abovementioned assumptions of the raters' true. For instance in paragraph 5, the writer described the advantages of studying individually according to the task. He used the statement of *'They say studying as a alone is more efficient.*' when he tried to imply that 'studying alone is more efficient.' The use of 'as' here is a negative transfer from the native language of the writer and not a convenient expression in English. Even though the NS raters knew Turkish in different competencies, some of them did not understand this phrase:

"That's a kind of weird phrase. I don't know what that is, but it's. Not a big problem. Just kill those two words - (as a alone). That's OK. Slightly weird way to say it." (NS 2)

"As a alone would not be proper, but I understand what he says." (NS 10).

While majority of the NS recognized that phrase and commented on it, most of the NNS either did not notice it or they did not commented on it. The statement of '*the walls shouldn't pass the voice"* in Paragraph 9 was also another example of the same concept. A NS who knew Turkish and a NNS realized this sentence and commented on it as follows:

"The walls shouldn't pass the voice, so obviously it doesn't know how to say. You know, sort of I forgot the English word for "ses yalıtımı", but doesn't you know able to express the opinion without actually knowing the word for that halls in another room shouldn't be noisy?" (NNS 7)

"He tries to say the walls should not transmit the voice. I could understand what he means because I'm Turkish." (NNS 10)

In light of this information, it can be concluded that NS raters can be more effective in detecting the errors deviate from the natural use of language given their mastery of the language's authentic usage. In addition, sharing the writers' native language facilitates a deeper understanding of their intended meaning and aids in distinguishing original utterances. In particular, to this study, NNS and NS who knows some Turkish were advantageous in identifying those errors and comprehending their underlying reasons for interlanguage transfer is useful to understand them.

Additionally, NS participants stated that their NNS colleagues might lack some information about the target language. Peculiar to English, they mentioned the most frequent mistakes were about collocations and prepositions, idioms, vocabulary choice and some parts of speech rules even though it did not create a remarkable difference and not a huge problem usually. In the interviews, most of the NS and NNS raters conveyed the same messages in that respect:

> *"Maybe easier for us to, you know, spot certain types of mistakes. Like, you know some grammar or some punctuation things or some. And some wording mistakes." (NS 9)*

> *"Native speakers may catch mistakes that we overlook. Issues that I might not deem significant could stand out to them. Additionally, they might have a better observer of the text's fluency." (NNS 9)*

One of the NS raters illustrated it with a vivid anecdote depending on her personal experience in her workplace:

> "*I think that there are some differences in how we respond to certain, especially to syntactical errors. For example, a couple weeks ago I was talking with an English teacher at my university. Turkish is her first language. She is a very good teacher. It was raining on that day. And then suddenly she said something like I'm really stressful when the weather gets like this". I am sure she learned that from her English teachers because it is the passive form of adjectives stressed interested, etc. I mean, Turkish people do not really distinguish between stressful and stressed. To me, as a native speaker, it's a syntactical error that, I really notice because that sounds wrong." (NS 2).*

These quotes suggest that some mistakes may appear insignificant to NNS or go unnoticed by them, even though they may know the correct form. However, these mistakes could significantly hinder the intended meaning, potentially more than NNS would realize. It is also revealed in this study that the most common mistakes made by Turkish people includes

collocations, prepositions, word form errors, -ed / -ing adjectives and idiomatic phrases according to the NS raters statements.

Finally, NNS mentioned the cultural preferences and their effect on the NS raters' attitudes. It was implied in the interviews that as a native member of the target language, NS raters normally behave more leniently and adopt a tolerant and understanding approach to the students' performance:

> *"Let's imagine for a moment that I am trying to teach Turkish to a non-native group. I think I can be more tolerant because I appreciate this effort due to emotional factors." (NNS 9).*

Conclusion

This research included 2 research questions. To answer these questions both qualitative and quantitative data were collected and analyzed. Quantitative data was derived from the assessment of 12 paragraphs written by students and analyzed using the Mann Whitney U test due to the non-parametric nature of the study. The results mainly implied insignificant differences between the scoring behaviors and rater consistency between these two groups apart from a few exceptions. The second research question was explored using qualitative data obtained through think-aloud sessions and interviews, analyzed using thematic analysis. The main aim of the second question was to gain a deeper insight of the NS and NNS raters' approach to writing assessment and identify any possible divergence in their strategies and tendencies. The results indicated important variances in their decision-making behaviors and their general assessment strategies were identified. In addition, some important points were revealed through interviews about the participants approach to fairness and reliability in assessment and scoring methods.

Discussion

Introduction

In this section, the findings presented in the previous section will be critically analyzed in relation to the relevant literature. The discussion will be organized into four sections. The first two of them correspond to the first research question and the rest to the second research question. These sections will draw upon important findings derived from descriptive statistics of the scorings, as well as insights obtained from think-aloud sessions and interviews. These discussions will be contextualized within the framework established in the literature review.

Rater Consistency in Native and Non-native Assessment

The primary objective of the first research question was to initially examine whether there are discrepancies in the NS and NNS assessment in terms of rater consistency and reliability, with the understanding that the more consistently a paragraph is assessed, the more reliable the scores. Two scorings methods were utilized to score 12 writings and as indicated by the findings, the scores awarded by the NS and NNS were not significantly different from each other for most of the paragraphs. This finding is contrary to previous studies, which have suggested that NS and NNS ratings varied in consistency (Kobayashi, 1992; Shi, 2001; Kim & Gennaro, 2012; Schaefer, 2008; Gonzalez, 2017). However, they agree with Lee (2009), Hill (1994) and Marefat & Heydari (2016) who claim that it is consistent. In fact, in the current study, NS scorings were found to be very slightly more consistent than NNS. However, the difference is unremarkable enough to ignore as in the study of Rao and Liu (2020). Correspondingly, it is put that obtaining full consistency is not possible in scoring writing since there is always room for subjective judgment in performance assessment. Therefore, ensuring up to 75% consistency among the raters is quite a sufficient rate to mention that it is ensured (Brown, 1990; Hamp-Lyons, 2003). As the literature suggests, concepts of consistency and reliability tend to go hand in hand. It can be argued that when results are consistent, both the test and the ratings are considered reliable (Hughes, 2003; Brown, 2004). Bringing together all the information about rater consistency, it may be accurate to state that NS and NNS raters in Turkey have developed a consistent, and consequently reliable, sense of judgment in writing assessment. They adhere to parallel writing assessment principles, suggesting that NS and NNS raters likely consider similar criteria in similar ways. This implies that NNS raters have developed a sufficient command of English comparable to NS. This finding contradicts Kobayashi

(1992), Hyland & Anan (2006) and Lee's (2009) assertions that NNS raters are less qualified than NS, as found in the studies conducted with Japanese, Korean and British participants. Based on this information, it can be assumed that writing assessment principles are standardized for university preparatory classes in Turkey, and both NS and NNS raters adhere to the same principles of writing assessment. This suggests that Turkey has adapted to the global writing assessment system, making it fairly reliable.

Severity in Native and Non-native Assessment

The other main concern of the first research question was to analyze severity in the NS and NNS raters and identify the differences, if any. The writing scores awarded by NS and NNS were compared statistically. Most of the paragraphs were scored in close grades by both of the groups. The findings of this study were in line with Connor-Linton (1995) and Lee's (2009) studies which identified no considerable difference in total scorings between NS and NNS in numerical state even tough they comment on the writings diversely and adopted varied strategies. This finding is contrary to previous studies which have suggested that NNS raters (Lascaratou, 1982; Song & Caruso, 1996; Kobayashi, 1992; Marefat & Heydari, 2015) or NS raters (Hill, 1994; Lee, 2016) score more severely than the other group in total scoring. When the details and several criteria has been considered, it was observed that the NS and NNS raters' approach to the scoring and their judgements might vary.

NS alluded to more specific grammar points (Connor - Linton, 1995) and provide more alternatives for correct lexis of the more natural way of the expressions (Kobayashi, 1992). It is implied in those studies that NS raters exhibit more confident behavior in scoring and interpretation for their language knowledge comes from their intuition and wider command of the language. The current study both in accords and contradicts with those results. First, they are inconsistent because both of the rater groups adopted quite confident behaviors with their professional knowledge and language competence unlike the studies suggesting that NNS were less eligible (Hill, 1994; Kobayashi, 1992), they are deficient in confidence in their competence (Hyland & Anan, 2006), and they are inferior to the NS raters in their performance and knowledge (Lee, 2009). There were few noticeable difference between them in terms of their qualifications in the current study. A possible explanation for these results might be that the language teaching education has been proved to be systematical, standardized and effective in Turkey. In other words, the raters receive a high-quality and internationally standard education on this topic. On the other hand, the same results can correspond with the idea that language intuition is helpful in language rating, which is the NS raters are more advantageous by birth. In the interviews

in this study, it was observed that some points required a background about the discrepancies with the natural use of the language. The statement of “the walls shouldn’t pass the voice” in the paragraph 9, can be an example for this case. More NS raters mentioned this statement and put forward that it is not the authentic or correct way to explain that situation. NNS however, were very few in number (3 raters) to spot it. It was named as negative transfer from Turkish by the NS. This term points to the negative influence of the learner’s native language on the language learning process, written expressions in this study. In other words, the learners are affected by their native language while they produce a text and the way they express themselves sounds more natural in their native language, not in the one they learn. Considering that grasping the authenticity in a language might be long and demanding process, it is understandable that the students need more time and exposure to build it. Correspondingly, NS raters can be more eligible in identifying and correcting negatively transferred expressions, especially when they know the learner’s native language, as in this study. However, this difference did not reflect on either the rater’s scorings or their interpretations in the current study due to the lower level of the writers, elementary and intermediate. In the interviews, some of the NS clearly stated when editing some higher level and more complex texts such as PhD theses. Thus, this study can be reconducted with a higher level group of writings and a convenient group of NS and NNS raters who are knowledgeable about the topic to identify the difference between their understandings.

In addition to abovementioned results, there is one more significant issues considering the raters general point of view. These results support the findings of the previous research which imply that NNS obtain a critical stance and made more negative comments than natives who adopted a more tolerant and understanding behavior (Kobayashi, 1992, Shi, 2001; Hyland & Anan, 2006) although few of them shows the opposite (Santos, 1988). This result may be explained by the fact that each study was conducted with a different group of non-native speakers (NNS), such as Koreans, Iranians, Japanese, Greeks or Indonesians, and NS raters such as British and American. Hence, it is important to account for social and cultural effects. It is sensible to remember that NS raters are on the natural speakers of the target language which positions them as the standard to which individuals from other cultures and nationalities aspire. It is quite understandable that the NS may exhibit leniency in their scorings. Peculiar to the current study, the NS demonstrated a more tolerant and understanding approach toward student mistakes as in Shi (2001), Hyland & Anan (2006) and Hughes & Lascaratou (1982), and their comments on them remained more tolerant and understanding while this leniency was

not reflected in the scores. Numerically, NS and NNS raters generally scored in parallel with each other. It can be concluded that the NNS raters have higher expectancy from the students of the same nationality with them. They were more strict in their intention to see the things they taught were correctly used as well as they were much less accepting against their mistakes. The shared nationality and cultural proximity between NNS raters and the students may result in a deeper understanding of the students' abilities. Conversely, a cultural gap between NS raters and students may lead to more lenient comments and approaches. Still, the lack of significant difference in numerical part of assessment implies that both groups ultimately arrive at similar judgements following different paths, which also points out the reliability of their assessments. These implications are also consistent with Connor-Linton (1995). The main motive behind this circumstance might be attributed to the comprehensive training both groups received in second language writing assessment. It can also be asserted that language assessment in second language learning has been systematically and scientifically developed, with internationally agreed-upon strategies in place.

Native and Non-native Approach to Assessment Criteria and Decision-Making Behaviors

The NS and NNS approaches to the assessment criteria were investigated through the first research question, but no significant differences were identified between the scores of NS and NNS raters. However, think-aloud sessions and interviews provided meaningful data about the divergence of NS and NNS perceptions of assessment criteria. That is to say, any meaningful numerical difference occurred in the current study for the final judgements of the two groups were in parallel with each other even though they emphasized separate aspects. Despite employing distinct strategies, both groups considered the same range of criteria, even when no predetermined criteria were provided for the impression scoring method. These findings broadly support the works of Connor-Linton (1995) and Rao & Liu (2020) whereas it contrasts with those of Marefat & Heydari (2015), who suggested significant divergence in NS and NNS scorings of important criteria. This result might be explained by the quality of language teacher education and its convenience with the international standards. It can be implied that English second language teachers, regardless of their mother tongue, developed a common intuition against writing assessment as a result of the education they received and their experience ensuring a reliable assessment approach for both groups.

Nevertheless, the interpretations and interviews conducted for the second research question revealed that NS and NNS raters exhibited different attitudes based on various assessment criteria. Previous studies noted that NS were concerned about global features such as content, organization, and coherence predominantly while NNS pondered linguistic features which are use of language, mechanics including spelling, and punctuation (Eckes, 2008; Kondo-Brown, 2002; Schaefer, 2008; Hyland & Anan, 2016). In the current study, on the contrary, the most important criteria was found to be content and organization for both of the groups, that is in keeping with Lee's study (2009). Most of both the NS and NNS chose content as the most important criteria in the interviews. Notwithstanding, content and organization are considered as global features which are linked to the meaning and communicative purpose of a written text (Hill, 1994; Song & Caruso, 1996; Lee, 2009). That might be the reason why they are usually intertwined and they were considered the most important factors in a written text. This finding also touches on the fact that both NS and NNS raters puts forward the meaning and development of ideas.

In terms of assessment, there was no significant differentiation between the two groups in terms of content and organization. NS were slightly more concerned with the content achievement, which was ascribed as being able to fulfill the task supporting the main arguments effectively with plenty of examples. This result is in line with previous studies in the literature (O'loughlin, 1992; Shi, 2001). In addition, it was observed that content was the very first criteria considered by the NS raters, which was followed by organization. On the other hand, NNS put organization forward in a minor manner as in Shi's study (2001). They mentioned more about organization and they had a pursuit of accurate use of language slightly more than the NS. They mentioned content right after they evaluate organization in general. For instance, in the think-aloud sessions, NS tended to read and grasp the gist of the writings and they made more comments about the quality of the ideas and examples given. In other words, they were in the pursuit of how well the ideas were developed and supported in the first place. However, besides that they also regarded the organization, looking out for topic, supporting, conclusion and linkers. Meanwhile, It was perceived that most NNS were apt to searched for the topic sentence and linkers primarily right before they mention the quality of the content. For example, the NNS raters predominantly referred to the topic sentence and supporting sentences whereas the NS raters sometimes ignored it or evaluated the organization altogether. Likewise, some NNS raters cut off points if the students used the "secondly" linker without introducing the supporting ideas with the "first" linker. However, NS evaluated this the same part more leniently and did not view this case as an issue since the meaning is clear. Moreover, the

NNS raters voted for the content and use of language when they were asked about the most important criteria even though they mainly consider organization while they score the writings contrary to their statements. These differences and contadiction could be attributed to the fact that NS raters act more comfortably about the meaning and evaluate the writings more flexibly while the NNS raters are likely to follow a set pattern and their comments were quite strict about the infraction of certain rules. The results of this study showed really close parallelism with some of the previous study which implies that content and organization is more important for both of the groups whereas NNS highlighted use of language and organization slightly more (Shi, 2001; Lee, 2009). Thus, this research contradicts with most of the previous study that upholds the idea that NNS are significantly more strict than NS in terms of accurate use of language (O'Loughlin, 1992; Rao & Liu, 2020; Marefat & Heydari, 2015, Lee, 2016, Hyland & Anan, 2006).

In addition to the contradictory outcome with the organization, another conflict were observed with the NNS assessment perception of mechanics. In the interviews, both the NS and NNS group of raters marked mechanics as the least important criteria. NS acted in accordance with their vote, concerning and referring mechanics less than the other criteria. However, NNS gave wide coverage to the mechanics at the scoring. Unlike their statements in the interviews, they reacted against the mistakes more than the NS, cut off points implying that mechanics is an important component of a good writing. They also complained about that it is usually neglected since it did not seem to be interrelated with the meaning. For this reason, the learners tended not to pay enough attention to mechanics. To promote the learners to do better with those criteria, the NNS raters confessed that they behave more severely and cut off more points to remark its importance.

Another factor that supports the same explanation is their approach to use of language. The general opinion about the raters approach to the assessment of use of language, implied by the outcomes of some studies (Lee, 2009; Hyland & Anan, 2006; Song & Caruso, 1996) suggests that NNS were too considerate and strict about the accuracy of grammar. Some of the NS raters, who also work with NNS Turkish instructors, agreed it in the interviews. However, only two of the paragraphs (9 and 11) received significantly different grades by the two groups of raters. Moreover, in contrast to earlier studies, NNS scored use of grammar more leniently. Nevertheless, as their thoughts were observed throughout the think-aloud sessions, NNS raters were more concerned about correct use of language and they were frustrated by basic mistakes. Not surprisingly, their interpretations were quite negative on this part. As an evidence, it can be reported that NNS

referred to subject - verb agreement, especially about am, is, and are much more times than NS raters who labelled them as basic mistakes that everyone does sometimes as mentioned in Hyland & Anan (2006). In addition, they tended to ignore some grammar mistakes if they did not influence the meaning. They did not referred to the grammar mistakes one by one. Instead, they mentioned them in general. On the contrary, NNS reacted fairly severely to those kind of mistakes and did not tolerate them implying that the students should have learned those fundamental rules by heart already. They were likely to note each one of them. The NNS raters' decision-making behaviors on considering error frequency also supported the NNS raters' tendency.

However, they also neglected the grammar mistakes if the meaning was not disrupted from time to time. Either way, it is important to emphasize that NNS raters were not influenced by their frustration while scoring and they graded akin to NS raters. This result may signify that both of the groups, NNS especially can act professionally and both of the groups regard content and organization as the most important criteria.

In this study, the NS and NNS raters' decision-making behaviors were also analyzed through certain assessment strategies inspired by Cumming et al (2002) & Sakyi (2000) and the results were found to be fairly consistent with theirs, assuming no considerable difference between NS and NNS occurred in majority of the assessment strategies. Only two of the strategies, "considering error frequency" and "considering spelling and punctuation" were alluded remarkably more by the NNS. NS, instead, took error gravity into account rather than the frequency. Most of the NNS raters were observed to count the errors one by one while scoring. Furthermore, some of them indicated that the rubrics should include instructions with the quantity of mistakes. It was claimed that the assessment would leave less room for the assessors' personal opinion and it would be more reliable in that way. These findings support earlier observations regarding the influence of organization, language usage, and mechanics on progressive decision-making behavior in writing assessment. This behavior entails assessors initially considering surface-level indicators such as topic sentences, supporting details, grammar, spelling, and coherence before evaluating the overall meaning (Cumming et al., 2002). To elaborate further, non-native speakers (NNS) tend to focus on discrete blocks of text first, then construct and assess meaning based on these blocks. Conversely, native speakers (NS) typically prioritize analyzing meaning from the outset, addressing smaller elements only if they impact overall coherence. This suggests a progression in assessment from linguistic to content-related criteria for NS, while NNS may follow an inductive approach, moving from specific details

to a general understanding. In comparing NS and NNS assessment approaches, it can be inferred that the former is deductive, while the latter is inductive, reflecting differing orientations toward specific-to-general and general-to-specific evaluation strategies. Nonetheless, these differences remained quite slight and affected neither the assessment process nor the scorings significantly.

Perceived Difficulties and in NS and NNS Writing Assessment

Both NS and NNS raters were interviewed and asked five questions. They all expressed similar difficulties, which provided a wider understanding of the second research question of this study. First, most of both NS and NNS clearly expressed that they found writing assessment stressful and demanding for it is systematically different from giving feedback to the students' paragraphs in a writing class. Trying to be fair in the assessment was confessed to be frustrating and also tiring by the raters. What is more, it was declared to be quite a time-consuming activity which might not be encouraging for most of the raters regardless of their native language. It was also mentioned by some of them that raters might not behave responsible enough to allocate from their time or they might simply not have enough time to do a detailed assessment. That might both create problems for the assessment reliability and quality and cause too much work and discomfort for the raters. However, this study also shed light on some possible solutions to these problems. The first one might be accepting that writing assessment is a subjective activity as it was mentioned in several sources (Barrett, 2001; Horowitz, 1991; Goulden, 1994, Hamp-Lyons, 1991; Vaughan, 1991; Kayapınar, 2014; Veigle, 2002; Casanave, 2004; Abedi, 2010; Rezaei & Lovorn, 2010) and expecting the ratings to always be exactly the same or very close may not be just and logical. As it is stated Hamp-Lyons conveyed, discrepancies up to 15% can be ignored and such an assessment can be considered reliable. As another possible solution, the raters mentioned that analytic assessment, which is used in most of the higher education institutions as the assessment method, can be helpful both to save time and energy and to give reliable and fair grades. This study made use of both impression scoring, a version of holistic assessment, and also an analytic assessment rubric. Thus, some strengths and weaknesses of the rubrics were observed. Firstly, the analytic assessment results proved to be undeniably reliable, offering a narrower range of differing ideas and scores due to the low level of standard deviation among the raters in this study. This result also aligns with findings from previous studies and sources (Hamp-Lyons, 1991; Weigle, 2002; Johnson & Svingby, 2007; Gonzalez, 2017). Ensuring reliability is an important factor for an assessment method for, as it was also discovered in this study, the weaker the

writings were, the wider were the score range. In other words, as the quality of a paragraph decreases, the variance in the interpretations increase. Analytic method and using a rubric can be solutions for the problems arising from assessment variance and subjectivity. However, it was repeatedly remarked that the raters bothered more than they were assisted when the analytic rubric was not designed comprehensively, covering any detail. Hence, they are beneficial only when they are designed well. To improve assessment conditions, rubrics should be developed by professionals according to institutional needs and periodically revised. Clear and comprehensive instructions are essential, eliminating the need for raters' personal opinions. An observation from this study is that minimum scores given through analytic assessment often exceed the writer's merit, evident in total scores and mentioned by raters in interviews. Combining holistic and analytic methods, despite personal biases, could mitigate this issue. Raters expressed the need for regular standardization training and emphasized the value of cross-marking for idea exchange and validation. However, the effectiveness of cross-marking depends on the rapport between partners, which may vary and lead to discomfort or stress. Standardization trainings are reported to be beneficial to improve assessment practice by the participants. On the other hand, a native speaker rater, who had previously worked in Cambridge, expressed concerns that raters assessing high-stakes exams like IELTS or TOEFL may struggle to adhere to the required conditions, regardless of the amount and the quality of training they receive. Based on the results and the insights provided by experienced participants in this study, it can be inferred that some raters possess a superior sense of judgment, intuition, and common sense in writing assessment. To enhance the reliability of assessment, establishing a core group comprising the most consistently reasonable raters could be advantageous, rather than relying on a large pool of raters. Such a core group is expected to reach consensus on assessment criteria and score writings more consistently and reliably. This group can even include NS and NNS raters at the same to enable a professional diversity.

Perceived Differences of Being a NS or NNS in Writing Assessment

In the light of the second research question, the interviews conducted with the both NS and NNS raters, revealed advantages and disadvantages of these groups as well as some of their strengths and weaknesses. NS, for example, were believed to have a better command of English as the natural users of the language. Not surprisingly it was expected that they have the mastery of idioms, phrases and correct word choice as the natural speakers of their language. As it was also discovered in the think-aloud sessions that NS can detect the negative transfer from the learners' native language. For some of the NNS

raters who share the same native language may overlook some incorrect use. They also may mot notice is there is a negative transfer from their main language for they share the native language. Hence, they may not be able to detect the inconveniency fot the expressions, depending on their native language may seem normal. As a matter of fact, 4 NNS can spot those kind of mistakes while the others did not pay attention at all. Plus, as some of the NS raters declared, sometimes their NS colleagues might seek for their advice with the convenience of the word choice, idioms or the most natural way to put something into words. In parallel with these findings, some previous research put forward that NNS were found not to be competent in English and these studies further discussed the NNS's the lack of qualification concluded in their inconvenience in writing assessment. One possible explanation to this result may be that NNS raters may lack information, or they neglect some points for this language is not intuitive to them. As it was discussed in this research, these are the reasons why NNS exhibited a lack of confidence in their assessment behaviors (Hill, 94; Lee, 2009). This may have led them to consider linguistic aspects such as use of language, spelling and mechanics instead of the global features such as content, coherence organization (Hyland & Anan, 2006). However, the findings of the current study do not support the previous research. The concordance between the NS and NNS raters' judgements and scorings, also stated in the NS raters' implied that NNS were quite well trained in terms of teaching and their second language (English in this case) competency. The differences between the two groups were quite slight.

NS participants made honest and clear interpretations on this subject depending on their own observation of their colleagues. They frankly stated that Turkish ESL instructors might have problems with collocations, prepositions, word-form errors, idiomatic phrases, some parts of speech errors, -ed / -ing adjectives the most. To be able to eliminate these problems, firstly, the institutions may include a NS rater for sharing information and supporting each other not only with the writing assessment but also with the other domains of teaching. Additionally, they can offer their teachers some training or courses where they can improve their command of English in a higher level. In this way, both the abovementioned problems can be minimized and for the teachers who teach or rate the lower levels, the problem of drifting apart from their language knowledge can be prevented. Conducting this study with the higher levels might carry the research one-step further. The difference between NS and NNS raters can be detected better with the higher levels. Besides the classes for teaching training, they might be provided with literature classes or some workshops with NS instructors, for instance. In addition, the teachers might change the levels they teach every year or semester to benefit from.

Finally, NNS raters who share their mother tongue with the learners can empathize better and understand the reasons behind certain mistakes and this might be the advantage in assessing writing. In this study, NS raters, who had some knowledge of Turkish, had lived in Turkey, and were familiar with the culture, also demonstrated an ability to understand why learners made specific errors. Through think-aloud sessions, it was observed that these NS raters were quite adept at identifying the reasons behind learners' mistakes. Thus, for NS raters, learning the students' mother tongue can be beneficial in understanding and assessing their writings more effectively, as well as enhancing their teaching performance in the classroom.

Conclusion

This study did not reveal any significant differences between the two groups overall, as their scores were largely parallel. This outcome may signify the standardization of the teacher education system. However, notable differences were observed in the raters' approaches to writing assessment, indicating that both groups have valuable perspectives and experiences to share. Writing assessment is a crucial aspect of language teaching, with foundational principles often established on an international scale. Nevertheless, incorporating diverse viewpoints can foster the development of both groups, leading to potential improvements in specific areas. These points will be further explored in the next chapter.

Chapter 5

Pedagogical Implications

Suggestions

Depending on its results, the current study can provide pedagogical implications that contribute to the application of writing assessment procedures. First and foremost, it acknowledges that writing assessment is a subjective process and that different evaluations or interpretations will always exist. Expecting each rater to agree on exactly the same points is unrealistic. In many institutions, two raters assess the same writing to ensure reliability, which can be stressful. Therefore, it is important to remind raters during courses or training to tolerate each other's ideas. This approach may help them be gentler and understanding, reducing the stress factor in the writing assessment process, which is not always appealing to raters.

In addition, this study has revealed that education and training in both writing assessment, and language teaching is the most important determinant criteria as Weigle corresponded with this idea stating that “Rater training is an issue that lies at the heart of both reliability and validity in ESL essay rating” (1994). The approaches of NS and NNS raters were observed and no significant difference were found. Also, during the think-aloud sessions, it was detected that the raters’ approach and steps they follow have fundamentally the same, with only minor variances. This result can signify that writing assessment education has been standardized and conducted according to international guidelines. One of the NNS raters clearly stated her idea as follows:

> *“We are very ambitious about this, and as a result, everyone is improving themselves significantly. I believe we have achieved a certain level of standardization in universities and similar institutions. This is why I think there is a noticeable difference. Additionally, there seems to be a consensus on this matter. (NNS 2)*

However, further studies comparing the behaviors of educated and uneducated NS and NNS raters could provide more insight into the impact of education on writing assessment, as all participants in this study were educated in their respective fields.

Mostly declared problems included some drawbacks writing assessment. Both the NS and NNS complained that it is a time-consuming, tiring and demanding process. Besides, it can also be stressful due to the raters’ effort to score the writings fairly. Three main solutions were prescribed for these problems mainly: standardization trainings, cross-

marking and detailed rubrics. All three are currently implemented in the institutions. However, to lessen the raters' ongoing discomfort, standardization trainings may be developed and for cross-marking the raters may be motivated to be understanding and supportive against each other, as it was mentioned above. Furthermore, the most repeated complaint was about the rubrics, which were thought to be more useful if they included more descriptors that are comprehensive. Developing rubrics to minimize the influence of raters' personal opinions is crucial. This approach not only prevents raters from exerting unnecessary effort but also ensures the reliability of writing assessments remains unaffected by subjective bias. Likewise, the raters might sometimes tend to complete the scoring as soon as possible to save their time and energy. When their answers of the first interview question about how they like writing assessment are considered, it can definitely be observed that this activity is not a very pleasant one for them. This might be another reason to lower the assessment quality and reliability. To prevent this, the institutions may think of some motivation to make it more of an advantageous activity such as offering a prize or they may provide more time to complete all their work within the bounds of the possibility of time so that the raters may divide the workload into some parts or share them with some partners. They need to be given motives to overcome their reluctance to invest time on it. Moreover only a core group consisting of them could be assigned with this work if it is feasible for as an important result of this study it is unveiled that some raters are just have more sensible judgements regardless of the trainings they received. If time and resources permit, it could also enhance reliability by reducing the range of interpretation variability. If not, they could be designated as the leader of the group to help with the inconsistencies having the final say. The institution might assign NS along with the NNS raters in those groups to share their different point of view.

Even though no significant difference revealed the NS and NNS pint of views to writing assessment, there are still some points that are worthy of touch upon to improve NS and NNS raters' abilities. NS could ensure to learn the language and culture of the group they teach. This would help them to understand the reasons of some common mistakes in writing assessment and possibly ease the teaching process as well. NNS raters, on the other hand, might be behaving unnecessarily negative on student while they only target to fix their manners. They might pay attention not to discourage students from making mistakes while learning either at class or through the assessment. If NNS raters' behavior of penalizing the student derives from any kind of anger or frustration, they should be reconsidering the underlying reason of their behavior.

The research shows a high degree of parallelism between the NS and NNS approach to writing assessment. Furthermore, they performed quite similarly in terms of

their command and competency of English. Thus, the belief that NS raters are more efficient for writing assessment was not supported by this particular study. On the contrary, how well the ELT teachers are educated has been discussed several times. A couple of NS who have worked with NNS raters in their institutions shared their supportive ideas as follows when the possible differences between the two groups:

> *"OK, I want to say the differences are not so remarkable. It doesn't make a big difference. Usually the non-native instructors that I work with have masters degrees that have studied for years. They have exhaustively studied grammar. They can usually name grammatical functions better than I can, but then I have a kind of magical sense. It's natural for me and I just understand it without trying to understand it. So it's two different kind of skills. I don't think one's better than the other. Sometimes NS instructors might be treated slightly differently like 'you are valuable because you're native'. I don't think that's always true. You guys are valuable because you work hard and you're good at your job." (NS 6).*

> *"Some NS are so poorly educated about their own language. I think the most important factor is education level. Honestly, I sometimes think like 'you are a 40 year old and you never learned it? Honestly, I'm sure that if I shared that with any of the Turkish colleagues who are, you know, Azeri, Belarussian, your colleagues who all learned English as a second language. And they're teaching English now. They would all laugh and make jokes about how poorly some teachers know their own language So I think that the educational level is much more important than whether someone is a native speaker or a non-native speaker." (NS 2)*

Nonetheless, several areas were identified that could benefit from improvement in the knowledge and performance of NNS raters. To start with, second language learners may have learn something wrong or may forget some points if they do not use it actively or expose the language regularly. Likewise, in case of the second language teachers, making mistakes does not mean that they are inefficient teachers. However, it is also understandable that second language teachers who to teach and assess the students language performance are expected to have a certain mastery of the language they teach. This might be the point where NS teachers have an advantage, as they recognize everything and know the correct versions that NNS teachers might forget, neglect, or simply not know one of the NS raters open heartedly shared an experience on this topic:

> *“A couple weeks ago I was talking with an English teacher at my university. Turkish is her first language and she's a very good teacher. On that that day that it was raining. She said something like ‘I'm really stressful when the weather gets like this’. The thing is that I'm sure she learned that from her English teachers because it’s passive form of adjectives like ‘stressed’, ‘interested, etc. I mean, most Turkish people don't really distinguish between stressful and stressed. To me, as a native speaker, it's a syntactical error that I really notice. But it's. As someone who's you know, learned to speak English with other Turkish. All these mistakes get fossilized and then they reinforcing each other because everybody got gets it wrong. So I think that NS teachers will respond to syntactical errors more than NNS teachers because we hear it and it sounds wrong.” (NS 5)*

NS teachers might mitigate these kinds of negative circumstances by being open to accepting that they may lack some information, which may deter them from performing their jobs thoroughly. They should also intend to continuously develop their command. According to the NS, the most frequent errors include collocations, prepositions, -ed/-ing adjectives, word form errors, and idiomatic phrases. These points might be specifically considered and the institutions might provide the teachers with some courses or training where they can foster their knowledge and English competency. Literature classes or workshops might serve as means of support in that respect. Having the same people teach at the lower level classes might be another factor that undermines their skills, but having them work at higher levels sometimes might be a good practice that keeps their language command alive. A simplified version of this study might be conducted with a higher level of students to detect if there is indeed a significant gap in English competency between NS and NNS teachers, which cannot be observed with lower levels in this study.

Putting aside all those details and focusing on the major finding, it can be concluded that NS and NNS raters do not necessarily vary in their judgements and scorings at the end of the scoring no matter how differently they approach to the topic. Main concern of writing assessment, as stated by raters from both groups, is about ‘how much one has to try to understand, which associates with content and meaning and at the end they arrived at similar conclusions in their evaluations Connor-Linton (1995) summarizes the core message of this study saying that: “They travelled different routes to arrive at similar destination”.

Conclusion

To conclude, the current study conducted in Turkey addresses the problems instructors encounter while assessing writing due to its subjective nature. Many members of the ELT community, working in various areas, either suffer from issues resulting from different interpretations or witness their colleagues experiencing these problems, as I have also experienced personally. This study also examines NS instructors who live and work in Turkey to determine whether they face similar issues and to find clues that might reduce subjectivity and bring about more consistent assessment practices. For this, it aimed at observing NS and NNS approach to writing assessment through a mixed method. Two groups of NS and NNS raters assessed 12 elementary and intermediate level writings in total. The data was collected through two writing assessment methods, impression scoring and analytic scoring. The analysis of scorings demonstrated very few significant differences ensuring the standards of education in writing assessment, which is the major finding of the current study. Some details observed in addition to this major finding, such as slightly more advantageous and disadvantageous sides of NS and NNS approach to writing assessment hopefully will contribute valuable insights to the current and future ELT community.

References

Abedi, J. (2010). Performance assessments for English language learners. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 199-214.

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

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

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

Barkaoui K. 2010. Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. TESOL Quarterly 44 (1): 31–57. https://doi.org/10.5054/tq.2010.214047

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

Baxter, A. (1997). Evaluating your students. London: Richmond Publications.

Bernard, H. R. (2002). Research methods in anthropology: Qualitative and quantitative approaches. Rowman.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific performance test. Language Testing, 12, 1–15.

Brown, J. D. (1996). Testing in language programs. Upper Saddle River: Prentice Hall Regents.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. TESOL quarterly, 32(4), 653-675.

Brown, H. D. (2003). Language assessment: Principles and Classroom Practices (1st ed.). Pearson Education.

Brown, H.D. and Abeywickrama, P. 2010. Language Assessment: Principles and Classroom Practices (2nd edition). White Plains, NY: Pearson Education.

Cambridge Dictionary. (n.d.). Native speaker. In dictionary.cambridge.org. Retrieved March 17, 2024, from https://dictionary.cambridge.org/dictionary/english/native-speaker

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied linguistics, 1(1), 1-47.

Carl, James. (1977). Judgements of Error Gravities, ELT Journal, Volume XXXI, Issue 2, January 1977, Pages 116–124, https://doi.org/10.1093/elt/XXXI.2.116

Casanave, C. P. (2004). Controversies in Second Language Writing: Dilemmas and Decisions in Research and Instruction University of Michigan Press.

Choi, Y. H. (2002). FACETS analysis of effects of rater training on secondary school English teachers scoring of English writing. Journal of the Applied Linguistics Association of Korea, 18(1), 257-292.

Connor-Linton, J. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. World Englishes, 14, 99–115.

Connor-Linton J. 1995. Looking behind the curtain: What do L2 composition ratings really mean? TESOL Quarterly 29 (4): 762–765. https://doi.org/10.2307/3588174

Creswell, J. W. (2021). *A concise introduction to mixed methods research*. SAGE publications.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.),Educational measurement (2nd ed., pp. 443–507). Washington, DC:American Council on Education.

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

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

Davies, A., Brown, A., Elder, C, Hill, K., Lumley, T., & McNamara, T. (1999). Dictionary of language testing. Cambridge: Cambridge University Press.

Dörnyei, Z. (2007). Research methods in applied linguistics. Oxford university press.

Ellis, R. (1986). Understanding second language acquisition. Oxford: Oxford University Press.

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

Engelhard Jr, G. (1994). Constructing Rater and Writing Task Banks for the Assessment of Written Composition.

Fachrurrazy, M. A. Definitions and Principles of Language Assessment.

Fayer, J. M. & Karshinski, E. (1987). Native and non-native judgements of intelligibility and irritation. Language Learning, 37, 313-326.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. *Educational measurement*, *2*(1), 237-270.

Fulcher, G. (2010). Practical language testing. London: Hodder Education.

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

González, E. F., Trejo, N. P., & Roux, R. (2017). Assessing EFL university students' writing: A study of score reliability. Revista Electrónica de Investigación Educativa, 19(2), 91-103. https://doi.org/10.24320/redie.2017.19.2.928

Goodman, G. S., & Carey, K. T. (2004). Chapter two: Critically situating validity and reliability. *Counterpoints*, *274*, 29-43.

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

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

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. New York: Longman.

Green, A., & Hawkey, R. (2012). Marking assessments: Rating scales and rubrics. *The Cambridge guide to second language assessment*, 299-306.

Hamilton, J., Lopes, M., McNamara, T., & Sheridan, E. (1993). Rating scales and native speaker performance on a communicatively oriented EAP test. Language Testing, 10(3), 337–353.

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

Hamp-Lyons, L. (2019). Assessing writing skills. In Dina Tsagari, Karin Vogt, Veronika Forehlich, Ildiko Csepes, Adrienn Fekete, Anthony Green, Liz Hamp Lyoons, Nicos Sifakis and Stefania Kordia (Eds), Handbook of assessment for language teachers (pp. 46-80). Erasmus+ (www.taleproject.eu)

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. Exploring the Dynamics of Second Language Writing (162-190). Cambridge University Press. https://doi.org/10.1017/CBO9781139524810.012

Harris, M., & McCann, P. (1994). Assessment. Oxford: Macmillan Heinemann.

Hayes, J. R., & Flower, L. S. (1981). Identifying the organization of the writing process. In L. W. Gregg, & E.R. Steinberg, Cognitive processes in writing (pp. 3-30). Hillsdale, NJ: Erlbaum.

Harmer, J. (2004). How to teach writing. Longman

Harmer, J. (2007). The Practice of English Language Teaching (Fourth Edition). Longman

Heaton, J.B. (1990). Writing English Language Tests: a practical guide for teachers of English as a second or foreign language. London. Longman.

Henning, G. (1996). Accounting for nonsystematic error in performance ratings. Language Testing, 13(1), 53–61.

Hill K. 1996. Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. Melbourne Papers in Language Testing 5 (1): 29–49.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. TESOL Quarterly, 37, 275-301.

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

Hughes A, Lascaratou C. 1982. Competing criteria for error gravity. ELT Journal 36 (3): 175–182. https://doi.org/10.1093/elt/36.3.175

Hughes, A. (2003). Testing for Language Teachers. Cambridge University Press.

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

Huot, B., O'Neill, P., & Moore, C. (2010). A usable past for writing assessment. College English, 72(5), 495-517.

Hyland K, Anan E. 2006. Teachers' perceptions of error: The effects of first language and experience. System 34 (4): 509–519. https://doi.org/10.1016/j.system.2006.09.001

Hyland, K. (2010). Metadiscourse: Mapping interactions in academic writing. *Nordic Journal of English Studies*, *9*(2), 125-143.

Hymes, D. (1972). On communicative competence. sociolinguistics, 269293, 269-293.

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

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

Kaplan, R.B. (1966). Cultural thought patterns in intercultural education. *Language Learning* 16, 1-20

Kasper, G. (1998). Analyzing verbal protocols. TESOL Quarterly, 32, 358-362.

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

Khalil, A. (1985). Communicative error evaluation: Native speaker's evaluation and interpretation of writer errors of oral EFL learners. TESOL Quarterly, 19(2), 335-35.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. Language Testing, 19(1), 3-31.

Kim, A.-Y., & di Gennaro, K. (2012). Scoring Behavior of Native vs. Non-native Speaker Raters of Writing Exams. Language Research 48.2, 319-342.

Kobayashi, T. 1992: Native and nonnative reactions to ESL compositions. TESOL Quarterly 26, 81–112.

Lee HK. 2009. Native and non-native rater behavior in grading Korean students' English essay. Asian Pacific Education Review 10 (3): 387–397. https://doi.org/10.1007/s12564-009-9030-3

Lee, K.R. (2016). Diversity Among NEST Raters: How do New and Experienced NESTs Evaluate Korean English Learners' Essays?. *Asia-Pacific Edu Res* 25 (4): 549–558. https://doi.org/10.1007/s40299-016-0281-6

Leki, I. (1992). *Understanding ESL writers*. NH: Heinemann Educational Books.

Longman. (2024). Native Speaker. In *Longman dictionary of contemporary English* (5th ed., pp. 304-305). Longman.

Machi, E. (1988). An exploratory study on essay-grading behavior of native-speaker and Japanese teachers of English. Paper presented at the 27th Annual Japan Association of College English Teachers Convention, Tokyo.

Macmillan Publishing Co, Inc; American Council on Education.

Marefat F, Heydari M. 2016. Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays. Assessing Writing 27: 24–36. https://doi.org/10.1016/j.asw.2015.10.001

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. Educational Measurement: Issues and Practice, 11(1), 3–9.

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

McDaniel, B. A. (1985). Ratings vs. equity in the evaluation of writing. Paper presented at the annual meeting of the Conference on College C,omposition and Communication, Minneapolis, MN. (ERIC Document Reproduction Service No. ED 260 459).

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103).

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

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

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. Journal of Applied Measurement, 4(4), 386-422.

Moskal, B. M., & Leydens, J. A. (2019). Scoring rubric development: Validity and reliability. *Practical assessment, research, and evaluation*, *7*(1), 10.

Norris, J. M. (1998). Designing second language performance assessments (No. 18). Natl Foreign Lg Resource Ctr.

O'loughlin, Kieran. (1992) Do English and ESL Teachers Rate Essays Differently? Melbourne Papers in Language Testing 1.2, 19-44.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. Language Testing, 19(2), 169–192.

O'Sullivan, B. (2000). Towards a model of performance in oral language testing. UK: University of Reading (Unpublished doctoral dissertation)

Ostler, S. (1987). English in parallels: A comparison of English and Arabic prose. In U Connor and R Kaplan (eds.) *Writing across languages: Analysis of L2 text*, Reading MA: Addison-Wesley.

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

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

Rao Z, Li X. 2017. Native and non-native teachers' perceptions of error gravity: The effects of cultural and educational factors. Asia-Pacific Education Researcher 26 (1-2): 51–59. https://doi. org/10.1007/s40299-017-0326-5

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. Assessing doi:10.1016/j.asw.2010.01.003

Ross, J. R. (1979).Where's English? In Fillmore, C.J., Kempler, D. & Wang, W. S-Y (Eds.). Individual differences in Language Ability and language behavior (127-163). Academic Press: New York.

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

Sandelowski, M. (2003). Tables or tableaux. *The challenges of writing and reading mixed methods studies Handbook of mixed methods in social and behavioral research*, 321-350.

Santos T. 1988. Professors' reactions to the writing of non-native-speaking students. TESOL Quarterly 22 (1): 69–90. https://doi.org/10.2307/3587062

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

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

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

Sheorey, R. (1985). Goof Gravity in ESL: Native vs Non-native perceptions. Paper presented at the 19th annual TESOL convention. New York.

Shi L. 2001. Native- and non-native-speaking EFL teachers' evaluation of Chinese students' English writing. Language Testing 18 (1): 303–325.

Shin, Y. (2010). A FACETS analysis of rater characteristics and rater bias in measuring L2 writing performance. English Language & Literature Teaching, 16(1), 123–142.

Slater, P. N. (1980). Remote sensing: optics and optical systems. *Reading*.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In Brindley, G. (Ed.), Studies in immigrant English language assessment, Volume 1 (pp. 159-189). Sydney: Macquarie University.

Song B, Caruso I. 1996. Do English and ESL faculty differ in evaluating the essays of Native English Speaking and ESL students? Journal of Second Language Writing 5 (2): 163–182. https://doi. org/10.1016/S1060-3743(96)90023-5

Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 15, 72-101.

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

Takashima, H. (1987). To what extent are non-native speakers qualified to correct free composition?-A case study. The British Journal of Language Teaching, 25(1), 43-48.

Think-aloud protocol by ozgur sahan. (2019, June 22). Youtube.

Vann, R., Lorenz, F., & Meyer, D. (1991). Error gravity: Faculty response to errors in the written discourse of non native speakers of English. In L. Hamp-Lyons (Ed.), Assessing second language writing in academic contexts (pp. 181-195). Norwood, NJ: Ablex.

Vaughan, C. (1991). Holistic assessment: what goes on in the raters' minds? In L. Hamp-Lyons (Ed.), Assessing second language writing in academic contexts (pp. 111-126). Norwood, NJ: Ablex.

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

Weigle, Sara Cushing. (2002). Assessing Writing . United Kingdom: (1st ed.) Cambridge University Press.

Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL writing: A pilot study. TESOL Quarterly, 37(2), 345-354.

Wolcott, W., & Legg, S. M. (1998). *An Overview of Writing Assessment: Theory, Research, and Practice*. National Council of Teachers of English, 1111 W. Kenyon Road, Urbana, IL 61801-1096 (Stock No. 34904: $18.95 members, $25.95 nonmembers)..

Yorkey, R. (1977). Practical EFL techniques for teaching Arabic speaking students. In J Alatis and R Crymes (eds.), The human factors in ESL . Washington: TESOL.

Zamel, V. (1983) The composing processes of advanced ESL students: six case-studies. TESOL Quarterly, 17: 165–87.

Zhenhui Rao & Fulan Liu (2020) An investigation into native and non-native teachers' assessment of Chinese students' English writing, Southern African Linguistics and Applied Language Studies, 38:2, 152-166, DOI: 10.2989/16073614.2020.1805776

APPENDIX-A: VOLUNTEER CONSENT FORM

Dear Colleague,

You are being asked to take part in a study that I am conducting within the scope of master's thesis under the guidance of Asst. Prof. Dr. Hatice Ergül in the Department of Foreign Languages Education at Hacettepe University. The main aim of my research is to compare the writing evaluation approaches of native and non-native English instructors and to inquire the differences between the two approaches. The necessary permission for my study was granted by Hacettepe University Ethics Commission.

Once you accetp to participate in this survey, you will take a questionnaire about your professional background,. You will also be asked to score 4 pieces of written texts thinking out loud. Finally you will have an interview about the details of your approach to scoring procedure. All in all, the survey should take 30 minutes approximately. All your responses, personal information and audio recordings will be kept strictly confidential. The findings will be used only for scientific purposes.

Participation is completely voluntary and it does not contain questions of personal discomfort of any kind. Nevertheless, if you feel uncomfortable with questions or you wish to stop answering for any other reason during participation, you may choose to stop participating at any time. In this case, the data you provide will not be used in the study.

If you have any questions concerning the research study or if you would like to be informed about the results, please do not hesitate to contact me.
Thank you in advance for your kind contribution.

"I am willing to participate in this survey."

Participant's:

Name, surname: ……..
Address: ……
Phone number: ……
Email: ……
Date: ……
Signature: ……

APPENDIX –B: STUDENT CONSENT FORM

Dear Participant,

I am a graduate student in the department of English Language Teaching at Hacettepe University. For my master thesis, I am conducting a research on native and non-native EFL instructors' evaluation process of the writing exams. With the purpose of investigating process, I would kindly ask your permission to take a copy of your writing exam to be scored by the instructors. The National Defence Univeristy approval has been received. If you have any questions concerning the research study or if you would like to be informed about the results, please do not hesitate to contact me.

Thank you in advance for your kind contribution.

Confidentiality

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

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

Student's:
Name, surname:
Class:
Signature:

APPENDIX –C: Personal Background Questionnaire

Dear Colleague,

You are being asked to take part in a study that I am conducting within the scope of master's thesis under the guidance of Asst. Prof. Dr. Hatice Ergül in the Department of Foreign Languages Education at Hacettepe University. The main aim of my research is to compare the writing evaluation approaches of native and non-native English instructors and to inquire the differences between the two approaches. The necessary permission for my study was granted by Hacettepe University Ethics Commission. It is important that you express your opinions sincerely for reliablity.

There are no risks related to participation in this study. Your participation will remain strictly confidential. Your name will not be attached to any of the data you provide. If you want to be informed about the findings, you can contect me.

Thank you very much for your participation and sincerity.

Ece Gürbüz
English Language Education Masters Student
Hacettepe University

Personal Information:

1. **Age:**

2. **Gender:**
   - ☐ female
   - ☐ male

3. **Native Language:**
   - ☐ English
   - ☐ Other

4. **Level of Education:**

   - ☐ Bachelor's degree The university and department that you graduated from:

     _____

   - ☐ Master's degree The university and program that you graduated from:

     _____

   - ☐ Doctorate (PhD) The university and program that you graduated from:

     _____

5. **Other Certificates and workshops:**

_____

6. **Did you have any trainig about writing evaluation?** (If yes, please give further information about it)
   - ☐ yes
   - ☐ no

_____

7. **The institution you work for:**
   - ☐ State university
   - ☐ Private institution

8. Year(s) of experience :

- 0 – 5
- 6 – 10
- 11 – 15
- 16 – 20
- 21 – 30

9. Do you think it's more reliable when essays are assessed by native speakers?

- yes
- no

10. **I go over the papers several times and check the previous papers I graded to be fair.**

- yes
- no

11. I sometimes rerate previous papers as I assess the others to create a balance.

- yes
- no

12. I do not hesitate to spend some time on considering the writing performance.

- yes
- no

13. I tend to finish rating as soon as I can.

- yes
- no

14. What is the most important criteria for you?

- content
- organization
- use of language (grammar)
- vocabulary
- mechanics

15. What is the least important criteria for you ?

- content
- organization
- use of language (grammar)
- vocabulary
- mechanics

APPENDIX –D: Analytic Rubric

Content	Organization	Use of Language	Use of Vocabulary	Mechanics and Spelling	Score
Text shows knowledge of the topic and gives details or examples to support main ideas. Text fully corresponds to task requirements. Communication is effective.	Organizational skills are present in the text. Ideas are coherent. Ideas flow smoothly. Main ideas and structure of text are clear and logically sequenced.	Text makes use of complex language structures effectively and maintains it. There are no errors of idioms, collocations and grammar in general. Facility in use of language is apparent.	Demonstrates sophisticated and broad use of vocabulary. Effective and appropriate use of idiomatic expressions and colloquialisms; shows awareness of connotations and their meaning.	Writing shows mastery of punctuation and spelling conventions. Capitalization and paragraphing errors and typos are not found.	5
Task is answered in its majority Sufficient development of main ideas. But information may be unnecessary, Some gaps may be found among information. More detail may be needed.	Adequately organized with the use of organizational patterns and connectors but sequencing of information is incomplete. Connection of main ideas may be lost but meaning is still understood.	Grammatical accuracy consistently maintained. Few errors of idioms, collocations and grammar in general. Complex sentences present minor errors.	Demonstrates sophisticated use of vocabulary. Good command of idiomatic expressions and colloquialisms. Minor errors in vocabulary use.	Writing shows occasional errors of punctuation and spelling conventions. Capitalization and paragraphing errors and typos are occasionally found.	4
Task is addressed adequately but information may be missing. Some details are used to support the main idea. Shows some knowledge of the main topic and limited development of main ideas.	Some organizational skills are present. Use of cohesive devices makes text clear. Occasional deficiencies can lead to "jumpiness" among information.	Some grammatical "slips" may be found. Grammatical errors such as verb tense, verb agreement, number, word order, articles, pronouns, and prepositions are found but they lead no misunderstanding.	Vocabulary accuracy is high though occasional errors may be found. Adequate and appropriate word/idiom choice and use. Some incorrect word choice does occur without impeding communication.	Writing shows few errors of punctuation and spelling conventions. Few capitalization and paragraphing errors and typos are found.	3
Task reveals little relevance to the topic. There are major gaps in information. Main ideas are not supported sufficiently. Pointless repetition of information.	Small pieces of text are linked with basic connectors. The level of cohesion is not satisfactory. The information is laid haphazardly. It causes the expressions seem non-fluent.	Frequent grammatical inaccuracies found. Frequent and basic errors of tense, agreement, number, word order, articles, pronouns, and prepositions are found. Understanding of ideas is seldom confusing.	Sufficient control of elementary vocabulary to express basic ideas. Repetition of vocabulary is frequent. Frequent misuse of word form use. Incorrect word/idiom choice, making communication confusing.	Writing shows frequent errors of punctuation and spelling. Capitalization and paragraphing errors and typos are frequently found. Meaning may be confusing.	2
Task presents limited relevance to main topic. Inadequate development of topic. Details are not given.	Groups of words connected with simple connectors such as "and", "but" or "because". Cohesion is almost absent. Connection among ideas is difficult to find making information confusing or misleading.	Most of the basic grammatical constructions are inaccurate. Major issues in simple sentences. Errors of agreement, number, word order, articles, pronouns, and prepositions frequently found. The text is difficult to understand.	Text has little knowledge of English vocabulary, idioms and word forms. Language sufficient for coping with simple survival needs. Inappropriate choice of word forms.	Almost all spelling is inaccurate and the text shows an ignorance of punctuation conventions . Text is dominated by capitalization and paragraphing errors and typos. Meaning is obscured.	1
Task does not reveal development topic. Totally inadequate answer to task. No details are given. Content is insufficient to assess.	Cohesion is totally absent. Writing is fragmented making communication impossible to obtain. Lack of structure in information leads to absence of organization. Content is insufficient to assess.	All language use is inaccurate. Meaning obscured. Content is insufficient to assess.	No apparent vocabulary use or vocabulary comprehension is present in text. Content is insufficient to assess.	All spelling is inaccurate and the text shows an ignorance of punctuation conventions . Text is dominated by capitalization and paragraphing errors and typos. Meaning is obscured. Content is insufficient to assess.	0

Adapted from: Council of Europe, 2009; CEFR, 2002; Jacobs et al., 1981; Weir, 1990; Gonzalez, 2017.

Content: It regards the content quality of the paragraph. It is expected that the required conditions in the question stem are met. If a certain conjunction, grammar, topic, word limit, etc. are covered in the question stem, the answer will be evaluated according to whether these are met or not.

Organization: It regards the logical and meaningful organization of the paragraph. The semantic integrity and fluidity of the writing answer will be evaluated in terms of word-subject repetition and appropriate conjunction use.

Use of Language: It regards language accuracy of the paragraph. The student's use of grammar rules he/she has learned within the scope of the course curriculum will be evaluated.

Vocabulary: The student's use of the words and word types he/she has learned within the scope of the course curriculum will be evaluated.

Mechanics: It regards the rules that the student has learned within the scope of the course curriculum, use of conjunctions and capital letters, punctuation marks, spelling mistakes, legibility, and paragraph formation.

APPENDIX –E: Ethics Committee Approval

Tarih: 15/12/2023 12:40
Sayı: E-66777842-300-00003255438

00003255438

T.C.
HACETTEPE ÜNİVERSİTESİ REKTÖRLÜĞÜ
Sosyal ve Beşeri Bilimler Araştırma Etik Kurulu

Sayı : E-66777842-300-00003255438 15/12/2023
Konu : Etik Kurulu İzni (Ece GÜRBÜZ)

EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

İlgi : 29.11.2023 tarihli ve E-51944218-399-00003224191 sayılı yazınız.

Enstitünüz Yabancı Diller Eğitimi Arabilim Dalı İngiliz Dili Eğitimi Tezli Yüksek Lisans Programı öğrencilerinden **Ece GÜRBÜZ**'ün, **Dr. Öğr. Üyesi Hatice ERGÜL** danışmanlığında yürüttüğü ***"Hazırlık Öğrencilerinin Yazma Becerilerinin Ana Dili İngilizce Olan ve Olmayan Okuyucuların Yaklaşımlarına Göre Değerlendirilmesi"*** başlıklı tez çalışması Üniversitemiz Sosyal ve Beşeri Bilimler Araştırma Etik Kurulunun **12 Aralık 2023** tarihinde yapmış olduğu toplantıda incelenmiş olup, etik açıdan uygun bulunmuştur.

Bilgilerinizi ve gereğini rica ederim.

Prof. Dr. İsmet KOÇ
Kurul Başkanı

Bu belge güvenli elektronik imza ile imzalanmıştır.

BelgeDoğrulama Kodu: CA67A947-8C5B-4F7E-94E7-C442739E9806
Adres:
E-posta: Elektronik Ağ: www.hacettepe.edu.tr
Telefon: Faks:
Kep:

Belge Doğrulama Adresi: https://www.turkiye.gov.tr/hu-ebys
Bilgi için: Burak CİHAN
Bilgisayar İşletmeni
Telefon: 03123051082

APPENDIX-F: *Declaration of Ethical Conduct*

I hereby declare that…

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

(DD) /(MM)/(YY)

(Signature)
Student's Name and Surname

APPENDIX-G: Thesis Originality Report

……/……./……

HACETTEPE UNIVERSITY

Graduate School of Educational Sciences

To The Department of Foreign Language Education

Thesis Title: Native And Non-Native Rater Approaches To Writing Evaluation In Preparatory Classes

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

Time Submitted	Page Count	Character Count	Date of Thesis Defense	Similarity Index	Submission ID
25 / 05 / 2024	115	141,911	31 /05 / 2024	13%	2387817740

Filtering options applied:

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

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

I respectfully submit this for approval.

Name Last name: Ece GÜRBÜZ

Student No.: N21130142

Department: Foreign Language Education

Program: English Language and Teaching

Signature

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

ADVISOR APPROVAL

APPROVED

(Title, Name Lastname, Signature)

APPENDIX-H: Yayımlama ve Fikrî Mülkiyet Hakları Beyanı

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

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

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

- Enstitü/ Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihinden itibaren 2 yıl ertelenmiştir. (1)
- Enstitü/Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren … ay ertelenmiştir. (2)
- Tezimle ilgili gizlilik kararı verilmiştir. (3)

…… /…… /……

(imza)

Öğrencinin Adı SOYADI

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

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

(2) Madde 6. 2. *Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç; imkânı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanın önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.*

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

Madde 7.2. *Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir*

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