JOURNAL OF LANGUAGE EDUCATION AND RESEARCH

Journal of Language Education and Research, 2023, 9 (2), 362-404

Research Article

Analysis of Raters' Scoring Behaviours in the Assessment of Writing

Yusuf Polat[*] **Nejla Gezmiş**[**]

ARTICLE INFO

Received: 16.07.2023
Revised form: 23.08.2023
Accepted: 27.08.2023
Doi: 10.31464/jlere.1328127

Keywords:
Foreign language teaching
writing
measurement and evaluation
scoring behaviours

ABSTRACT

This study aims to obtain descriptive data on the scoring behaviours of language teachers in the assessment of writing. In this study, the data related to the demographic information of 73 teachers, the scores they assigned to a particular written text, and the discourse describing their scoring process were obtained through semi-structured interviews. Data analysis revealed that the scoring behaviours of the teachers varied greatly and most of them did not use a scale during scoring. The ones who used scales, on the other hand, created their own scales that included different sections with either the same or different scoring attributes, which indicated that they were indecisive and not necessarily unified in their decisions. Another finding indicated that the teachers focused mostly on the formal dimension of the product, and only a limited number of them distinguished error types and considered the balance between the error type and the points. The results of the study also demonstrated that there was a significant relation between the scores of the raters and their gender, as well as between their scores and their institution.

Acknowledgments

We would like to thank the teachers who spent their time and effort in scoring the written products for their contribution to the data collection of this study.

Statement of Publication Ethics

The ethics committee approval has been obtained for the current study: Kırıkkale University Social and Human Sciences Research Ethics Committee with decision no. 09 dated 18/10/2022.

Authors' Contribution Rate

The contribution rate of the authors is equal.

Conflict of Interest

This study has no conflict of interest.

Reference

Polat, Y., & Gezmiş, Y. (2023). Analysis of raters' scoring behaviours in the assessment of writing. *Journal of Language Education and Research, 9*(2), 362-404.

[*] Prof. Dr., ORCID ID: https://orcid.org/0000-0001- 9341-6643, Kırıkkale University, Department of Translation and Interpretation in French, polatyus@gmail.com

[**] Asst. Prof. Dr., ORCID ID: https://orcid.org/0000-0003-4909-1460, Kırıkkale University, Department of Translation and Interpretation in English, nejlagezmis@gmail.com

ISSN: 2149-5602

Introduction

Writing is the process of producing a written discourse for communication. Based on the communication scheme, writing is defined as the process in which an addresser sends a written/visual message formed by using a common code (English, French, Turkish, etc.) to an addressee through a contact (paper, screen, etc.) in a certain context. As one of the main four skills in foreign language teaching, writing merits an important place, particularly in terms of measurement and evaluation, along with other aspects.

There are two main approaches, namely direct and indirect, in the assessment of writing. The direct approach, which considers all aspects of the relevant skill (Coombe, False, and Hubley, 2007), evaluates the learner's ability to communicate in a written language. In such assessment, learners are expected to produce content, organize their ideas, use appropriate and accurate vocabulary, and apply grammatical and syntactic knowledge within a certain period in the classroom through traditional measurement tools. The fact that such assessment can be applied in any setting within a limited time makes it easy to apply, however, it also has some negative aspects such as putting time pressure on learners and not allowing for full implementation of the stages of the writing process as preparation, planning, drafting, a nd reviewing. In indirect assessment, on the other hand, the focus is on the grammatical accuracy of a language rather than its communication function (Coombe, Fals,e and Hubley, 2007). Accordingly, it is possible to evaluate the correct use of components such as punctuation, spelling, and grammar at the sentence level by using measurement tools such as multiple-choice tests. In recent years, process-oriented assessment, where the measurement and evaluation of writing are spread throughout the process, has been used in the evaluation of the process, and the development of learners during this process is monitored. Although this type of assessment bears more reliable results, it is time-consuming and extends over a long period (Brown, 1989).

O'Malley and Pierce (1996) argue that the measurement and evaluation process for writing requires giving students writing tasks on various topics and evaluating their product using its message, clarity, and mechanical aspects such as spelling and punctuation. In other words, rather than a one-dimensional approach, writing assessment requires a multidimensional approach that takes into account the complex structure of the writing process. However, one of the most important problems of teaching writing in both first language and foreign language in Türkiye is the inability to conduct a practical and consistent measurement and evaluation (Karatay, 2011). Since writing involves not only vocabulary, grammar, spelling, and punctuation of language but also more complex processes, the approaches to be used in its measurement should also include more comprehensive and complex processes of writing. In their study conducted with 97 English teachers through survey and interview techniques in Türkiye, Kalay, and Büyükkarcı (2020) found that apart from using the traditional assessment tools, including multiple choice questions, completion, true-false or question-answer, the teachers also utilized process-oriented tools, as composition writing, project assignment, peer evaluation, portfolio, and standard writing tests, all of which refer to more complex processes.

The effective assessment of writing, be it traditional or process-oriented, depends on the reliability, validity, and practicality of the measurement instrument (Weigle, 2007). Reliability in the assessment of writing is generally lower than in the other language areas and skills (Calp, 2013). Mousavi (2002) also emphasizes the importance of measurement and evaluation instruments in the assessment of writing. Determining the writing task and developing appropriate instructions are among the key factors affecting the validity of measurement and evaluation. Hence, it is paramount to provide a task that will clearly reveal the direction of the writing that is intended to be measured and to add an instruction that clearly explains the expected behaviours of learners throughout the process. Instructions have important functions not only for learners but also for raters. To enable a consistent and reliable measurement of the written product and its evaluation, task instructions need to be taken into consideration by raters to determine whether the learners have written in line with these instructions and to evaluate their writing with an objective perspective.

Considering various types of written products with differing requirements, it is necessary to determine to what extent those requirements are met during the assessment process. Thus, another important topic regarding the assessment of writing in the literature involves the scoring method performed. There are three fundamental scoring methods: analytical, holistic, and primary trait scoring (Cooper, 1984; Perkins, 1983; Stiggins and Bridgeford, 1983; Weigle, 2007; Wiseman, 2012; Zorbaz, 2013). The scale used in analytical scoring, namely the rubric, allows separate evaluations for each component of writing. It provides detailed information on learners' strengths and weaknesses; however time-consuming and challenging for raters since it requires a detailed examination of the written product and identification of deficiencies. In contrast, the holistic scale is more time and energy-efficient it enables raters to make general judgments about the written product. Primary trait scoring is task-specific and evaluates performance according to the specific features of the discourse type. Therefore, the scales used in this type of assessment include items related to the characteristics of discourse and text type. The advantage of primary trait scoring is that it provides the testing of learners' knowledge and skills related to the characteristics of the genre, and it facilitates the detection of learners' shortcomings specific to the discourse type. Yet, it is less frequently used than the other two scales as developing it is more demanding and requires more time and expertise.

The abundance of studies on scoring methods and the diversity of their results are noteworthy. Wiseman (2012) found in his study, in which 60 compositions written by EFL students were scored through analytical and holistic scales by five teachers who had special training in scoring, that analytical scoring revealed more detailed and accurate differences among learners. Sakyi (2000), on the other hand, argued that the raters who scored holistically focused more on language structures and content. While Polat (2003) and Turgut (1990) claim that analytical scoring is more reliable because it provides conformity between raters, Oruç (1999) argues that inconsistencies between raters decrease with the usage of holistic scoring. However, in their study with 10 raters, Han and Huang (2017) asserted that there were no significant differences in terms of scores or reliability between the two scoring methods, however, the raters preferred the holistic

scoring scale regardless of the advantages and disadvantages of both types. However, there are conflicting studies yielding opposite results. For example, in their study in the context of EFL teaching in Yemen, Ghalib and Hattami (2015) concluded that the differences between raters were greater in holistic scoring, and they claimed that analytical scoring provides a more reliable and consistent measurement. Similarly, Asassfeh (2021) proposed that the scores assigned by 48 teachers by using the holistic scales were higher compared to the scores they assigned by using analytical scales, and that the scores of the raters decreased in holistic scoring since they focused on the details.

Although the studies on scoring methods vary and reveal different results, it is obvious that the use of scales brings objectivity to the assessment of writing. A study conducted by Crusan, Plakans, and Gebril (2016) with 701 educators revealed that 80% of the participants used a scale and that nearly half of the participants believed that using a scale was effective in making sense of students' points. A scoring scale is an important tool not only for teachers but also for learners. While assigning a writing task as homework or as a measurement tool to learners, it is of great importance to determine the criteria that will be used in the evaluation of that task and to present them to learners on a scale. Seviour (2015) emphasizes that scales should be open and accessible to students, and they should be informed about what is expected of them, how the written product will be evaluated, and, what qualities an acceptable written product should have. On the other hand, Thomas (2020) argues that when developing or choosing a scale, there are some important factors to be considered such as, what will be measured by this scale, for what purpose this scale will be used, how long it will take to administer it, and whether training will be necessary for the users, all of which will contribute to the reliability and validity of that scale.

Behaviours of Raters

Another effective factor in the assessment of writing is raters. Smith (1993) stated that the knowledge and the measurement technique of raters affect the reliability of their measurement. This is not surprising because a written product is complex and the professional experience, education, and views of the person scoring it play an important role on their scoring. Therefore, raters should receive qualified training on measurement and evaluation. Köksal (2004) claims that the teachers in Türkiye do not receive adequate training on the assessment of writing and they have a tendency to assess their students by using general assessment ways presented in the curriculum. Indeed, when examining the curriculum of the institutions that train teachers in foreign language teaching, it was observed that general training on measurement and evaluation is provided, but there is a lack of training specifically focusing on the assessment of each skill while individual skills are separately taught. Similarly, in the international literature, there are studies claiming that teacher training is insufficient as far as measurement and evaluation are concerned (Brown and Bailey, 2008; Mertler, 2009; Popham, 2009; Weigle, 2007). Therefore, it can be suggested that the level of literacy in terms of assessment of writing is low among teachers. Likewise, in their study with 350 teachers, Mede and Atay (2017) concluded that

the teachers perceived themselves as inadequate in measuring and evaluating production and reception skills within integrated skills.

However, for reliable measurement and evaluation to be carried out, the people who carry out the process, namely the raters - teachers in the school environment - need to be highly competent in the field of measurement and evaluation. Crusan (2010) emphasizes that teachers should know the differences between formative and summative assessment, have the ability to write instructions that will provide the necessary guidelines for presenting data required for different purposes, know the priorities of the scales used, and comprehend the importance of assessment. Equally, Weigle (2007) underlines that teachers should have the skills in organizing, managing, and scoring writing activities in order to assess writing. In conclusion, raters are expected to develop a measurement and evaluation instrument and score the written product with the help of this instrument.

As stated by Baker (2016), the scoring behaviours of raters can be influenced by factors such as personality, education, and their desire to appreciate learners' effort and ability to understand what they are trying to convey. However, raters should evaluate written products with a valid, objective, fair, clear, systematic, criterion-based, and reliable measurement method and process. While many studies underline the need for raters to use a scale during the assessment of writing, Lumley (2002) points out that the way the raters used the prescribed scale in his research was quite inconsistent. Similarly, many studies using the 'Rasch Model' (Du & Wright, 1997; Du, Wright & Brown, 1996; Engelhard, 1994; Lunz, Wright & Linacre, 1990) proved that the raters' scores differed even if they had received the same training and used the same scale. The fact that there are differences in raters' priorities, expectations, and the dimensions they focus on the written product at hand during the measurement and evaluation processes is among the factors underlying this situation.

The errors of raters can affect the validity and reliability of measurement and evaluation (Erman Aslanoğlu and Şata, 2021). Raters assign lower/higher scores due to various characteristics of learners such as gender, race, experience, expertise or because of the handwriting, paper layout, and the argument of the written product indicates the rater effect in writing assessment. While Gyagenda and Engelhard (2009) detected that the raters did not exhibit different scoring behaviours based on the students' gender, Engelhard and Myford (2003) unearthed that the raters scored differently based on the students' gender, race, and language, in which they were most successful. Johnson and Lim (2009) established that there was a slight difference in the scores between the raters who were native speakers and those who were non-native speakers. Erman Aslanoğlu and Şata (2021) deduced that the raters took into account the overall academic achievement levels of students, but not their gender. They also found that the teachers working in state schools exhibited different scoring behaviours from the ones working in private schools.

The assessment of the written product which is composed at the end of complex processes is also a multifaceted and problematic process. The first of these problems is the excessive workload caused by the process. The workload of teachers at school, the excessive number of students in the classroom, and the long time and effort required for the evaluation of written products make it difficult to conduct a comprehensive

assessment. Secondly, it is quite difficult to use completely objective criteria in the assessment of writing due to the absence of having a correct, precise and complete answer for the expected written product. Thirdly, raters may focus not only on the dimensions included in the scale but also on their own internal evaluation criteria, despite being presented with a scale (Li & Huang, 2022). While written products should be evaluated in an objective way, it is claimed that teachers score written products based on their own expertise and impressions of the paper in question (Çetin, 2002). During the impression-based evaluation process, noticeable mechanical errors are usually marked and the layout of the text or paper is taken into consideration. In a case study conducted by interviewing 12 teachers in Kayseri, Göçer (2011) collected the teachers' views on writing assessment. This study reported that the teachers carried out a collective assessment, most of them used different measurement tools in addition to writing compositions, and they experienced difficulties concerning time and application during the assessment process. Besides, it was found that they did not use a common scale, and they centered their attention around formal qualities involving plan, tidiness, layout of the paper, spelling, and punctuation during the assessment of writing.

To sum up, the factors affecting objectivity in writing assessment can be listed as the measurement process that includes scoring, raters and scoring methods, the characteristics of learners such as gender, age, race, ethnicity, social class, learning environment, the elements related to writing task itself and its instructions, and also the dimensions in the scale (Gyagenda and Engelhard, 2009). There are also numerous studies (Calp, 2013; Cole, Haley & Muenz, 1997; Hamp-Lyons, 2002; White, 1994) showing that the assessment of writing poses problems. Among these problems are the measurement and evaluation methods (Beck, Llosa, Black and Anderson, 2018; Cooper, 1984; Han and Huang, 2017; Şeker, 2018; Tokur Üner and Aşılıoğlu, 2022; Wilson et all., 2016), validity and reliability of measurement instruments (Brown, Glasswell and Harland, 2004; O'Neill, 2011), and the consistency and reliability of the individuals who performed the measurement and evaluation (Erman Aslanoğlu and Şata, 2021; Gyagenda and Engelhard, 2009; Lumley, 2002; Smith, 1993; Wind and Engelhard, 2012; Zhang, 2016).

Most of the studies on writing assessments focus on the characteristics of the measurement instruments, measurement methods, and the types of scales used during the assessment process. Clearly, in these studies, the participants are provided with a scale to apply and they are guided in the process of assessment. It is also evident that the textbooks in Türkiye contain scales. Additionally, it is assumed in the studies that teachers use scales in light of the guidance in their ordinary writing assessment process. Although there are many studies investigating teachers' assessment techniques and methods, the number of studies describing teachers' scoring behaviours is limited. Furthermore, no research on how teachers will evaluate without guidance has been found in the literature. Therefore, this study aims to examine the scoring behaviours of foreign language teachers in writing assessments. In line with this aim, answers to the following questions and sub-questions are sought within the scope of the study:

1) What kind of scoring behaviours do foreign language teachers exhibit when scoring a written product in the assessment of writing?

a) Is there a statistically significant relation between the scores of the foreign language teachers and their demographic characteristics?

b) What kind of marking behaviours do foreign language teachers exhibit when scoring a written product in the assessment of writing?

c) How do foreign language teachers approach students' errors when scoring a written product in the assessment of writing?

d) What dimensions of the written product do foreign language teachers focus on when scoring a written product in the assessment of writing?

2) How do foreign language teachers describe their scoring process when scoring a written product in the assessment of writing?

Methodology

This study aimed to identify the scoring behaviours of foreign language teachers by examining and comparing the papers they scored. It is also aimed to identify their ways of scoring by analyzing their descriptions of the scoring process. For this purpose, the teachers were first asked to score a randomly selected composition which was written in English and French by B1-graded students studying at the Department of Translation and Interpreting, and then to describe their scoring process.

Participants

A total of 73 foreign language teachers voluntarily participated as raters in the study. The age of the participants, most of whom were female, varied between 31-40 years old. Most of the participants were teachers of English. In addition, most of them had training in pedagogical formation and had over 15 years of experience. Furthermore, most of them were working at the university level at state institutions. The detailed data regarding the demographic characteristics of the participants are collectively presented in Table 1.

Data Collection Instruments

The data of the study were obtained through a semi-structured interview technique. The interview form consisted of three sections. The first section included the items about demographic characteristics of the participants such as gender, age, graduated program, professional experience, current working level, and institution. In the second section, the participants were asked to score the given composition out of 100, and they were asked to briefly describe their scoring process in the third section.

Table 1. Demographic Characteristics of the Participants

Variable	Category	Number	Percentage (%)
Gender	Male	15	20,5
	Female	58	79,5
Age	Between 20-30	11	15,1

	Between 31-40	30	41,1
	Between 41-and 50	24	32,9
	51 and over	8	11,0
Graduation	ELT	48	65,8
	Other	25	34,2
Pedagogical Formation	Yes	71	97,3
	No	2	2,7
Experience	1-5 years	6	8,2
	6-10 years	15	20,5
	11-15 years	14	19,2
	16-20 years	17	23,3
	21 years and over	21	28,8
Level	Elementary	6	8,2
	Secondary	7	9,6
	High School	20	27,4
	University	35	47,9
	More than one	5	6,8
Institution	State	62	84,9
	Private	11	15,1

Publication Ethics

Ethics committee approval for the interview form was obtained from the Kırıkkale University Social and Human Sciences Research Ethics Committee with the decision no. 09 dated 18/10/2022.

Data Analysis

Document analysis and descriptive analysis were used in data analysis. In the first stage of the analysis, the focus was on the demographic characteristics of the raters, and the data obtained from the first section of the interview form were interpreted through document analysis. In the second stage, observations were made regarding the second section of the data collection instrument, and the way the participants scored the composition was identified through document analysis and descriptive analysis, and the scoring behaviours of the teachers were determined. These behaviours were first coded and placed into the appropriate category in the thematic framework based on the research questions. In the third stage, the third section of the interview form was utilized. For this purpose, the participants' descriptions of their own evaluation processes were read and the scoring behaviours they expressed were determined by the researchers. Likewise, these stated behaviours of the raters were coded and placed into the abovementioned framework. Shortly, three datasets including teachers' demographic characteristics, scoring behaviours and, discourses on their own scoring behaviours were obtained in the study. Comparative analysis of the second and third datasets are made based on the first dataset. Moreover, examining the second and third datasets enables us to determine whether there is consistency between teachers' discourses and actions as well as whether there is a

relationship between teachers' scoring behaviours and their age, gender, and experience. The findings are presented in numbers and percentages in tables.

The main limitation of the study is that only composition writing was used as a technique in the assessment of writing. Since composition writing requires not only students' grammatical knowledge but also their syntactic, semantic, pragmatic, and textlinguistic knowledge, it was thought that all kinds of writing skills could be measured in this way most efficiently. As the aim of the study was to examine the scoring behaviours of the participants, one composition in English and one in French written by two students were used instead of the ones produced by different students. Thus, the factor of individual differences between students was eliminated from the study. Although different languages are taught as foreign languages in Türkiye, the study is limited to the teachers of English and French due to the principle of practicality as they are the most commonly taught foreign languages.

Findings

Findings Related to the Rater's Behaviours

When the scores of the participants were examined, it was seen that the lowest score was 40 and the highest score was 100 out of 100 points. The most important finding about the scores is that teachers appointed different scores for the same writing product. The detailed findings about the scores are presented in Table 2.

Table 2. Findings About the Scores of the Participants

Scores	Number	Percentage (%)
Between 81-100	41	56,2
Between 61-80	20	27,4
Between 40-60	7	9,6
At least 90	1	1,4
Total Score	1	1,4
No Score	2	2,7
Total	73	100

Two out of 73 teachers were not included in the analysis because they did not specify any scores. Thus, the analysis was accomplished through the scores of 71 teachers. As seen in Table 3, the data did not exhibit a normal distribution since the Skewness and Kurtosis coefficients were not between $-1<p>1$, and the Kolmogorov test result was $p<0.05$.

Table 3. The findings on the coefficient of Skewness, Kurtosis, and Kolmogorov Smirnov

Number of the Participants	Mean of the Scores	Skewness Coefficient	Kurtosis Coefficient	Kolmogorov Smirnov Coefficient
71	82,89	-1,336	1,767	.000

Since the data did not exhibit a normal distribution, the Mann-Whitney test was used to see if there was a difference between the scores given by the raters according to their gender. At the end of the analysis, a significant difference was observed between the scores in favor of female raters (see Table 4). It is understood that female raters gave higher scores than male raters because their mean and mean rank were higher.

Table 4. The Findings on the Analysis of the Scores According to Gender

Category	Number	Mean	Mean Rank	Total Rank	U	P
Female	57	84,67	38,89	2216,50	234,500	.017
Male	14	75,64	24,25	339,50		

As identified before, the participating teachers were working at state or private institutions. The Mann-Whitney test was used to identify whether there was a difference between the scores of the raters in terms of the institution they worked at, as the data did not show a normal distribution. As seen in Table 5, there was a significant difference between the scores given by the raters according to the variable of the institution they were working. It is concluded that teachers working at state schools gave higher scores than those working at private schools since their mean and mean rank were higher.

Table 5. The Findings on the Analysis of the Scores According to the Working Institutions

Category	Number	Mean	Mean Rank	Total Rank	U	P
State	61	85,23	38,87	2371,00	130,000	.004
Private	10	68,60	18,50	185,00		

To find out whether the scores of the teachers varied according to their professional experience, the data were analyzed using the Kruskall Walls test because of the five categories in the theme of experience. As a result of the analysis, it became clear that there were no significant differences between the scores given by the teachers and their years of professional experience (see Table 6).

Table 6. The Findings on the Analysis of the Scores According to the Professional Experience

Category	Number	Mean	Mean Rank	sd	Mean	P
1-5 years	6	73,17	29,25			
6-10 years	14	77,43	24,29			
11-15 years	13	84,77	39,00	4	7,688	.104
16-20 years	17	87,00	42,91			
21 years and over	21	84,81	38,29			

To examine whether the scores of the teachers varied according to the level they were working, the data were analyzed using the Kruskall Walls test since the working level contains five different categories. As seen in Table 7, there was a significant difference ($p=0.036<0.05$) between the scores of the teachers according to their working level. An

examination of their averages revealed that the teachers working at multiple levels and teachers working at university gave lower scores than the others.

Table 7. The Findings on the Analysis of the Scores According to the Working Level

Category	Number	Mean	Mean Rank	sd	Mean	P
Elementary	6	86,33	37,50			
Secondary	6	85,83	37,42			
High School	19	86,11	40,32	4	10,265	.036
University	35	83,97	37,14			
More than one	5	55,40	8,10			

When it came to the sub-questions b, c, and d of the study, another analysis was conducted on the participants' marking behaviours on the paper, their approach to the type of error, their preferences for interaction with the student and the errors they focused on. For this purpose, the answers to the following questions were sought: a) whether they did marking or not, b) if yes, how many markings they did, c) whether they identified the types of errors or not, d) whether they warned the students or not, e) whether they corrected the errors or not, f) whether they used a coding system during marking, g) whether they focused on linguistic mistakes (grammatical, spelling, punctuation, lexical, syntactic errors) or not, h) whether they took the types of discourse into consideration during scoring or not, i) whether they paid attention to the correctness of the content or not, j) whether they used a scale or not. The detailed findings of this analysis are shown in Table 7. This analysis mostly reveals that the raters usually corrected errors on the paper and they usually focused on grammatical and expression errors.

Table 8. The Findings on the Behaviours of the Participants

Category of the Behaviours	Number	Percentage (%)
Only marked	30	41,1
Identified error types	8	11,1
Warned the students	11	15,1
Corrected errors	47	64,4
Adopted a coding system	2	2,7
Focused on grammatical errors	64	87,7
Focused on spelling errors	35	47,9
Focused on punctuation errors	26	35,6
Focused on lexical errors	31	42,5
Focused on expression errors	32	43,8
Paid attention to the features of discourse type	12	16,4
Took the correctness of the content into consideration	5	6,8
Used a scale	10	13,7

Findings Related to the Rater's Discourse

When the participants' description of their scoring process is analyzed in order to find an answer to the second research question, two main dimensions emerge the usage of a scale and the direction of their focus. As shown in detail in Table 9, the former

dimension includes the findings about the number of raters who acknowledged using the scale, as well as who developed a scale on the paper, the points they allocated to each error or section in the scale, and the number of sections used in the scale. Prominently, most of the raters drew their own scale on paper themselves and evaluated the paper accordingly. However, there were significant differences in the implementation of the scale among the users (See Table 9). The most prominent behavioural difference in this regard is the fact that the numbers of the sections in the developed scale were not the same. Another striking point is that the participants anticipated the same or different points for each section in the scale they developed.

Table 9. The Findings on the Participants' Discourse About the Usage of a Scale

Scale Preference		Number	Percentage (%)
Expressed that they used a scale		4	5,5
Drew a scale		26	35,6
The number of sections in the scale	2	1	1,4
	3	3	4,1
	4	9	12,3
	5	7	9,6
	6	1	1,4
Allocated the same points to each section in the scale		16	21,9
Allocated different points to each section in the scale		10	13,7
Allocated no points in the scale		2	2,7
Stated how many points each error corresponded to		3	4,1

Analysis of the raters' focus on their own descriptions reveals various preferences. Accordingly, the raters indicated that they took various aspects such as unity, coherence, cohesion, relevance to the subject, content, language usage, consistency, comprehensibility, grammar, punctuation, spelling, vocabulary, number of words, time, target audience, purpose, features of the genre, sections of text, organization, evidence/examples, paper layout, handwriting, creativity, style, fluency, attention grabbing, way of thinking, participation in the class, age, planning, and level of the students into considerations during evaluation. Having analyzed the number and frequency of these aspects, it was displayed that only one participant took the variables related to target audience, participation in class, age, nationality, and creativity into consideration, which corresponded to a ratio of 1,4 for each. While the target audience refers to the addressee of the written pro and participation in class means the in-class performance of the students, age indicates whether the producer of the written product is a child or an adult and nationality indicates the producer's familiarity with the language used in the composing process. As can be seen in Table 10, the variables that the participants considered relatively of high importance were grammar, vocabulary, meaning and the organization of the text while they also attributed attention to paper layout, fluency, way of thinking, relevance to the subject, unity, evidence/examples, features of the genre, sections of text, level of the students, language usage, coherence, spelling, content and punctuation.

Table 10. The Findings on the Analysis of the Raters' Focus

Direction of the Focus	Number	Percentage (%)
Target audience	1	1,4
Participation in the classroom	1	1,4
Age	1	1,4
Nationality	1	1,4
Creativity	1	1,4
Style	2	2,7
Consistency	2	2,7
Time	2	2,7
Attention Grabbing	2	2,7
Planning	2	2,7
Handwriting	3	4,1
Cohesion	4	5,5
Number of the words	4	5,5
Purpose	4	5,5
Layout	8	11,0
Fluency	8	11,0
Way of thinking	8	11,0
Relevance to topic	9	12,3
Unity	10	13,7
Evidence/Examples	11	15,1
Features of text type	12	16,4
Sections of text	14	19,2
Level	14	19,2
Language usage	15	20,5
Cohesion	16	21,9
Spelling	16	21,9
Content	20	27,4
Punctuation	20	27,4
Organization	22	31,5
Meaning	24	32,9
Vocabulary	35	47,9
Grammar	68	93,2

To sum up, the data obtained revealed a number of findings related to the research questions. The first set of findings concerns the actions that teachers performed on the paper they scored. As a result, no common preferences were observed among the teachers regarding marking approaches. While some preferred marking approaches such as underlining or circling, some others preferred to mark by specifying the error type.

The second set of findings pertains to the scores given by the teachers. Accordingly, the participants were assigned significantly different scores for the same paper scores given to which same paper ranging from 40 to 100. The third set of findings, which is related to the usage of a scale in the scoring process, revealed that they were used as well as those who did not. The fourth set of findings is about whether there was a difference among the scores according to raters' gender, institution, experience and grade.

The final set of findings involves the focus of the participants in the measurement and evaluation process, which concluded that a significant number of the participants focused on the grammatical aspect of the written product, respectively followed by vocabulary, meaning, organisation, punctuation, and content of the output, all of which could be listed under the heading of grammatical aspect of writing. Non-grammatical aspects such as cohesion, language use, and arts of the text were considered by relatively fewer participants.

Discussion

The findings that were significantly different in the scoring preferences of the participants need to be addressed in several ways. First of all, differences in scoring can be regarded as natural findings since the participants were expected to score without any given standard scales. There are numerous studies (Bachman, 2004; Engelhard and Myford, 2003; Hunter and Docherty, 2011; Liu, 2022; Şeker, 2018) concluding that the scoring of the raters was different when they were asked to screen with the help of the same standard scale. However, the fact that differences among the scores were in a wide range of 40 to 100 makes the scores appointed by the raters questionable in terms of reliability. The low reliability of the raters' scores in this study is consistent with the results of a study conducted by Gyagenda and Engelhard (2009), in which 20 raters who were provided with training on scoring, scored 366 compositions using a scale and the reliability coefficient was found to be low. The findings obtained from the same study and our findings are noteworthy in that they proved that teachers had different preferences in scoring irrespective of whether they had any training on scoring or not, and whether they were asked to use a standard scale or not. The difference in scoring was evident in all sections of the scale in that study, whereas in our study it was observed that the teachers proposed the same or different points for each section of the scale only when they developed their own scale. Allocating different points to each section of the scale may stem from either the raters or the fact that the variable cannot be expressed as an absolute value. While the raters' behaviours such as giving little or much importance and showing necessary attention to the section during the evaluation process may lead to differences arising from the rater, the fact that the evaluated dimension does not have a single and ideal answer due to its nature may also lead to differences depending on the evaluated dimension. On the other hand, although some dimensions such as grammar can be relatively expressed as more absolute value, the fact that they were scored differently suggests that scoring differences cannot be eliminated in the evaluation process despite all efforts. Şeker (2018) endorsed similar findings regarding the behaviours of the raters/teachers in the assessment of writing in the study conducted with three English teachers working at a school in Türkiye. Writing was evaluated through a standard scale. 75 of the compositions, which were written by the students as a writing exam, were selected by equally dividing them as low, medium and high level, and three teachers were asked to score 25 of these compositions with the help of the same scale in three days. Then, the teachers were asked to score the other 25 of compositions by discussing them together and this process was recorded. Three weeks later, the teachers were again asked to individually score the last remaining, 25, of the composition with the help of the same

scale. The analysis of the first scoring verified that three teachers scored the same papers differently even if they used the same standard scale. It was understood that the teachers made different judgements on different aspects of the written product such as grammatical accuracy, lexical accuracy, syntactic accuracy, organization, and mechanical dimensions. The statistical analysis of the data unearthed that the points that those three teachers assigned to the items in the scale were not compatible with each other. When the recordings of the scoring were analyzed, it was seen that the teachers had hesitations about their decisions on scoring on the first day, and that they relaxed, exchanged their ideas and showed their expertise in different dimensions on the second day. For instance, while the argument of one teacher on grammatical accuracy was taken as a basis, the argument of another teacher on organization was taken as a basis during the discussion period. In addition, it was also found that while the scoring process lasted longer on the first day, it was getting shorter day by day. It became obvious that the teachers considered the scoring fairer at the end of this process and they shared the responsibility as they did not issue a score on their own. When the data of the scoring were analyzed three weeks later, it was found that the teachers scored the dimension of grammatical accuracy, lexical accuracy, organization and mechanical aspects similarly. Hence, this study suggests that the teachers scored the same written product in different ways despite using the same standard scale, and their focus differed during the evaluation process; for example, some of the teachers focused on accuracy while others focused on fluency or structure. One of the notable findings of Şeker's study is that the participants showed different reactions at different stages of the discussion process, which lasted for three days in total. To exemplify, while they were mostly silent and exhibited hesitant behaviours on the first day of the scoring discussion sessions, they engaged in discussions confidently and made more confident decisions without hesitation on the following days. In addition, the teachers also discussed the scale they used and identified its deficiencies. It was also deduced that the differences among the scores of the teachers decreased in the subsequent individual evaluation, and they used the experience and knowledge that they had gained through discussing in their own individual scoring sessions. Therefore, this study suggests that cooperation and discussion with other stakeholders during the scoring process provide benefits for the raters and bring consistency to the evaluation process. Compared with the findings of Şeker's study (2018), the findings obtained from the current study are similar in some aspects such as the teachers' attempts to prepare the scales, which show inconsistency, and the differences in the scales developed by themselves.

The finding that suggested significant differences between male and female raters in scoring in favor of female raters is inconsistent with some research data. For instance, in a study conducted by Peterson, Childs and Kennedy (2004), 108 teachers in the first language teaching in Canada were asked to score the narrative and argumentative compositions produced by two female and two male students. The results of the study unveiled that there were no consistent findings about the difference between the scores of raters in terms of their gender and that there was no significant difference among the scores depending on the gender of the composition writer. However, there are other studies (Gyagenda and Engelhard, 2009) supporting that there were differences in scoring based

on gender. For example, a study conducted by Gyagenda and Engelhard (2009) reports that the differences between the scores of male and female students were significant in favor of females. The researchers assert that the reason for this difference might be a prevailing belief that female students were more successful in writing or that the teachers focused more on the written product generated by male students in order to improve their writing skills. Yet, this study did not take the gender of the 20 trained raters into account in the analysis.

The finding that the number of sections and the points per section varied among some participants is important as it suggests that the teachers perceived and explicated the same written product in different ways. The differences in preferences of the participants, most of whom had pedagogical formation training, are in parallel with the findings of the study by Wang et al. (2017), which yields that there were different opinions on various points between the experts delivering training on scoring and the raters, and also among the participant raters. For instance, there were disagreements involving the selection of the most difficult composition to score, the sections leading to the errors of raters, the focus of the written product, the reception of the text, and the organisation of the ideas between the experts delivering training on scoring and the raters. These disagreements are important in understanding that if the scale is not clear and understandable enough for the raters, it can lead to different results even if the same scale is used. However, it is suggested that no matter how valid and reliable the scale is, the rater's knowledge, experience, and attention are more decisive in perceiving and using the scale.

The findings that the participants focused on extremely diverse dimensions of the written product during the evaluation process and that they issued different terms for the same dimensions assume that the raters had different experiences and education in this aspect. As an example, while the dimension titled 'grammar' in the scales needed to be a general heading to the other three as syntax, spelling and punctuation, it was treated as another dimension along with them. Similarly, some of the dimensions titled as language usage, coherence, cohesion, unity, consistency and fluency are unclear and this raises the question of whether the rater differentiates, for example, between consistency, unity and coherence or between coherence and cohesion. Furthermore, the concept of language usage should include the dimensions listed above, however, that raters' evaluation of it as a separate dimension shows that they had different perspectives and perceptions on this issue. In line with the findings of the research by Wang et all (2017), which suggests more experiential studies should be conducted, this finding highlights the fact that the distinctions regarding the different dimensions of the written product should be more clearly put forward and strongly emphasized in the training on scoring. This idea is reinforced by the findings of Rahayu's (2020) study conducted with 56 ESL teachers in Indonesia in the assessment of writing. In that study, the teachers were asked to answer questions about their assessment methods and techniques and to score two narrative compositions with the help of an analytical scale presented to them. The questions in the questionnaire consisted of four sections that were related to the teachers' knowledge about the assessment of writing, the effectiveness of scoring accuracy, the efficiency of their choices, and their perceptions of the assessment of writing. The analysis of the data

obtained through the questionnaire yielded that the teachers' knowledge about the assessment of writing, the effectiveness of scoring accuracy, the efficiency of their choices, and their perceptions in the assessment of writing did not guarantee their success in scoring. In other words, the teachers' responses in the questionnaire did not show consistency with their scoring behaviours. Even though teachers' knowledge about the assessment of writing, the efficiency of their choices, and the increase of their perceptions in the assessment of writing negatively affected their scoring, the increase in the effectiveness of scoring accuracy positively affected their scoring. The result of the study reinforced that the effectiveness of teachers in scoring has an impact on the quality of teachers during the assessment process.

Another finding that the raters focused more on grammar and the formal dimension of the written product is in line with the findings commonly found in the literature. In their study on the perceptions of the raters in the assessment of writing through an integrated approach, Weigle and Montee (2012) revealed that while the raters attached different importance to the formal dimension of the written product, they also exhibited different attitudes towards the reception technique used by the students in the writing process.

Conclusion and Implementation

This study which aims to identify the scoring behaviours of foreign language teachers in writing assessment reveals that the raters, who are foreign language teachers participating in this study as a study group, exhibited very different behaviours when scoring the written product. Despite the limitations, this study has postulated important results regarding the assessment of writing, which is accepted to be problematic in the literature. In this context, it is obvious that the question of whether foreign language teachers use a criterion while scoring a written product cannot be answered entirely in a positive way. It is seen that the number of those who use a criterion among the participants is limited. To generalize this result, it would be useful to conduct more studies with different study groups to determine the diversity of the behaviours exhibited by the raters while scoring and also whether they use criteria or not.

The research question about determining which dimensions of the product the teachers focused on during scoring was inquired and the finding that the teachers focused more on the formal dimension is deemed to be in accordance with the literature findings. However, many studies in the literature recommend that all dimensions of the written product should be considered in the assessment of writing. In response to the third research question, the study asserted that a limited number of raters discriminated error types and established equivalence between the severity of errors and the points appointed. At this point, it is supposed that teachers' knowledge about the assessment of writing is limited. Therefore, teacher training institutions must include courses related to the assessment of writing in their curriculum. Additionally, it is crucial for those who are currently working as teachers to upskill themselves and embrace every opportunity to improve their knowledge in this field through various courses or seminars.

Finally, the research question regarding whether there are differences in scoring behaviours of raters in terms of their age, gender, and experience yields results indicating

that female raters gave higher scores to the written product than their male counterparts, the raters working at state institutions gave higher scores in contrast to those working at private schools, the raters working at more than one level and at the university level gave higher scores than others, whereas there was no significant relationship between the experience and the scores of the raters. Inquiryinton the existence of these differences in various groups and determining the underlying reasons for these differences could add a new perspective to the studies on the rater effect in the assessment of writing. Doubtlessly, trying to identify and eliminate the factors that cause the rater effect makes the assessment process of writing more reliable.

References

AbuSeileek, Asassfeh, S. M. (2021). Holistic vs. analytic scoring between expository and narrative genres: Does the assessment type matter? *International Journal of Linguistics, Literature and Translation*, *4*(1), 215-220. https://doi.org/10.32996/ijllt.2021.4.1.21

Bachman, L. F. (2004). *Statistical analyses for language assessment book*. Cambridge University Press.

Baker, K. M. (2016). Peer review as a strategy for improving students' writing process. *Active Learning in Higher Education*, *17*(3), 179–192. https://doi.org/10.1177/146978741665479

Beck, S. W., Llosa, L., Black, K., & Anderson, A. T. G., (2018). From assessing to teaching writing: What teachers prioritize. *Assessing Writing*, *37*, 68-77. https://doi.org/10.1016/j.asw.2018.03.003

Brown, J. D. (1989). Manoa writing placement examination. *Manoa Writing Board Technical Report*, 5.

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing*, *25*(3), 349–383. https://doi.org/10.1177/02655322080901

Brown, G. T. L., Glasswell, K., & Harland, D. (2008). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*, 105-121. https://doi.org/10.1016/j.asw.2004.07.001

Calp, M. (2013). Serbest ve yaratıcı yazma tekniğine göre oluşturulan kompozisyonların yazılı anlatımın niteliği ve puanlama tekniği açısından karşılaştırılması. *Turkish Studies: International Periodical Fr the Languages, Literature and History of Turkish or Turkic*, *8*(9), 879-898. https://doi.org/10.7827/turkishstudies/5340

CECR (2018). *Cadre européen commun de référence pour les langues: Apprendre, enseigner, évaluer*.. Retrieved on May 15, 2023 from www.coe.int/lang-cecr

Cole, J. C., Haley, K. A., & Muenz, T. A. (1997). Written expression reviewed. *Research in the Schools*, *4*(1), 17–34.

Coombe, C. A., Folse, K. S., & Hubley, N. J. (2007). *A practical guide to assessing English language learners*. University of Michigan.

Cooper, C. G. (1997). Holistic evaluation of writing. In C. R. Cooper, & L. Odell (Eds.). *Evaluating writing* (pp. 3-33). National Council of Teachers of English.

Cooper, P. L. (1984). The assessment of writing ability: A review of research. *Educational Testing Service*. https://doi.org/10.1002/j.2330-8516.1984.tb00052.x

Crusan, D. (2010). *Assessment in the second language writing classroom*. University of Michigan.

Crusan, D. Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs and practices. *Assessing Writing*, *28*, 43-56. https://doi.org/10.1016/j.asw.2006.03.001.

Çetin,B. (2002). *Kompozisyon tipi sınavlarda kompozisyonun biçimsel özelliklerinden kestirilen puanların anahtarla ve genel izlenimle puanlanmasından elde edilen puanlarla ilişkisi* [*The relation between scores predicted from structural features of an essay and scores based on scoring key and overall impression, in essay type examination*] Unpublished MA Thesis, Hacettepe University, Ankara.

Du, Y., & Wright, B. D. (1997). Measuring student writing abilities in a large-scale writing assessment. In M. Wilson, Jr G. Engelhard, & K. Draney (Eds.). *Objective measurement: Theory into practice* (pp. 1-24). Abex.

Du, Y., Wright, B. D., & Brown, W. L. (1996). Differential facet functioning detection in direct writing assessment. *In Annual Conference of the American Educational Research Association* 8-12 Nisan 1996 (pp. 1-21). ERIC.

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

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

Enginarlar, H. (1991). A quantitatıve and qualitative comparison of three techniques of grading ESL/EFL essays. *Journal of Human Sciences*, *10*(1), 23-45. https://www.j-humansciences.com/ojs/index.php.IJHS/issue/view/27.pdf

Erman Aslanoğlu, A., & Şata, M. (2021). Examining the differential rater functioning in the process of assessing writing skills of middle school 7th grade students. *Participatory Educational Research*, *8*(4), 239-252. https://doi.org/10.17275/per.21.88.8.4

Ghalib, T. K., & Hattami, A. A. (2015). Holistic versus analytic evaluation of EFL writing: A case study. *English Language Teaching*, *8*(7), 225-236. https://doi.org/10.5539/elt.v8n7p225

Göçer, A. (2011). Öğrencilerin yazılı anlatım çalışmalarının Türkçe öğretmenlerince değerlendirilmesi üzerine. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi*, *30*(2), 71-97. https://doi.org/10.7822/egt34

Gyagenda, I. S., & Engelhard Jr, G. (2009). Using classical and modern measurement theories to explore rater, domain and gender influences on student writing ability. *Journal of Applied Measurement*, *10*(3), 225-246. https://d1wqtxts1xzle7.cloudfront.net/32596450/Gyagenda_Engelhard-libre.pdf

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, 8, 5-16. https://doi.org/10.1016/S1075-2935(02)00029-6

Han, T., & Huang, J. (2017). Examining the impact of scoring methods on the institutional EFL writing assessment: A Turkish perspective. *PASAA: Journal of Language Teaching and Learning in Thailand*, *53*, 112-147. https://files. eric.ed.gov.tr/fulltext/EJ1153666.pdf

Hunter, K., & Docherty, P. (2011). Reducing variation in the assessment of student writing. *Assessment and Evaluation in Higher Education*, *36*(1), 109-124. https://doi.org/10.1080/02602930903215842

Jakobson, R. (1960). Linguistics and poetics. In T. A. Sebeok (Eds.). *Style in language* (pp. 350-377). Mass. MIT.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505. https://doi.org/10.1177/0265532209340186

Kalay, S., & Büyükkarcı, K. (2020). English language teachers' views on teaching and assessment of writing skills. *SDU International Journal of Educational Studies*, *7*(2), 262-286. https://doi.org/10.33710/sduijes.710062

Karatay, H. (2011). Süreç temelli yazma modelleri: Planlı yazma ve değerlendirme. In M. Özbay (Eds.). *Yazma eğitimi* (pp. 21-43). Pegem Akademi.

Köksal, D. (2004). Assessing teacher' testing skills in ELT and enhancing their professional development through distance learning on the net. *Turkish Online Journal of Distance Education*, *5*(1), 1- 11. https://dergipark.org.tr/en/pub/tojde/issue/16931/176755

Liu, L. (2022). Scoring judgment of pre-service EFL teachers: Does writing proficiency play a role?. *The Asia-Pacific Education Researcher*, *31*(3), 333-343. https://doi.org/10.1007/s40299-021-00575-9

Li, J., & Huang, J. (2022). The impact of essay organization and overall quality on the holistic scoring of EFL writing: Perspectives from classroom English teachers and national writing raters. *Assessing Writing*, *51*, 1-15. https://doi.org/10.1016/j.asw.2021.100604

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters?. *Language Testing*, *19*(3), 246-276. https://doi.org/10.1191/0265532202lt230oa

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331-345. https://doi.org/10.1207/s15324818ame0304_3

Mede, E., & Atay, D. (2017). English language teachers' assessment literacy: The Turkish context. *Dil Dergisi*, *168*(1), 43-60. https://dergipark.org.tr/tr/pub/dilder/issue/47674/602254

Mertler, C. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools*, *12*(1), 101–113. https://doi.org/10.1177/1365480209105575

Mousavi, S. A. (2002). *An encyclopedic dictionary of language testing* (3rd ed.). Tung Hua.

O'Malley, J. M., & Pierce, L. V. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. Addison-Wesley.

O'Neill, P. (2011). Reframing reliability for writing assessment. *Journal of Writing Assessment*, *4*(1), 1-15. https://escholarship.org/uc/item/6w87j2wp

Oruç, N. (1999). *Evaluating the reliability of two grading systems for writing assessment at Anadolu University preparatory school*. Unpublished MA Thesis, Bilkent University, Ankara.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures and objective tests to evaluate ESL writing ability. *TESOL Quarterly*, *17*(4), 651-671. https://doi.org/10.2307/3586618

Peterson, S., Childs, R., & Kennedy, K. (2004). Written feedback and scoring of sixth-grade girls' and boys' narrative and persuasive writing. *Assessing Writing*, *9*(2), 160-180. https://doi.org/10.1016/j.asw.2004.07.002

Polat, M. (2003). *A study on developing a writing assessment profile for English preparatory program of Anadolu University School of Foreign Languages*. Unpublished MA Thesis, Anadolu University, Eskişehir.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice*, *48*, 4–11. https://doi.org/10.1080/00405840802577536

Rahayu, E. Y. (2020). The anonymous teachers' factors of assessing paragraph writing. *Journal of English for Academic and Specific Purposes*, *3*(1), 1-19. https://doi.org/10.18860/jeasp.v3i1.9208

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Eds.). *Fairness and validation in language assessment: Selected papers from the 19th language testing research colloquium* (pp. 129-151). Cambridge University.

Seviour, M. (2015). Assessing academic writing on a pre-sessional EAP course: Designing assessment which supports learning. *Journal of English for Academic Purposes*, *18*, 84-89. https://doi.org/10.1016/j.jeap.2015.03.007

Smith, W. L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M. M. Williamson & B. A. Huot (Eds.). *Validating holistic scoring for writing assessment* (pp. 142-205). Hampton.

Stiggins, R. J., & Bridgeford, N. J. (1983). An analysis of published tests of writing proficiency. *Educational Measurement: Issues and Practices*, *2*(1), 6-19. https://doi.org/10.1111/j.1745-3992.1983.tb00679.x

Şeker, M. (2018). Intervention in teachers' differential scoring judgments in assessing L2 writing through communities of assessment practice. *Studies in Education Evaluation*, *59*, 209-217. https://doi.org/10.1016/j.stueduc.2018.08.003

Thomas, N. (2020). Idea sharing: Are analytic assessment scales more appropriate than holistic assessment scales for L2 writing and speaking? *PASAA: Journal of Language Teaching and Learning in Thailand*, *59*(1), 236-251. https://files.eric.ed.gov.tr/fulltext/EJ1239980.pdf

Tokur Üner, B., & Aşılıoğlu, B. (2022). İngilizce öğretiminde ölçme ve değerlendirme sürecine ilişkin öğretmen görüşleri. *EKEV Akademi Dergisi*, *89*, 25-50. https://dergipark.org.tr/en/pub/sosekev/issue/71371/1147452

Turgut, M. F. (1990). *Eğitimde ölçme ve değerlendirme metotları*. Saydam.

Wang, J., Engelhard Jr, G., Raczynskia, K., Song, T., & Wolfec, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing*, *33*, 36-47. https://doi.org/10.1016/j.asw.2017.03.003

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, *16*(3), 194-209. https://doi.org/10.1016/j.jslw.2007.07.004

Weigle, S. C., & Montee, M. (2012). Raters' perceptions of textual borrowing in integrated writing tasks. *Studies in Writing*, *27*, 117–152. https://doi.org/10.1163/9789004248489_007

White, E. (1994). Issues and problems in writing assessment. *Assessing Writing*, *1*, 11–27. https://doi.org/10.1016/1075-2935(94)90003-5

Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing*, *27*, 11-23. https://doi.org/10.1016/j.asw.2015.06.003

Wind, S. A., & Engelhard Jr, G. (2012). Examining rating quality in writing assessment: Rater agreement, error, and accuracy. *Journal of Applied Measurement*, *13*(4), 1-15. https://d1wqtxts1xzle7.cloudfront.net/31697482/SW_GE_2012-libre.pdf (d1wqtxts1xzle7.cloudfront.net)

Wiseman, C. S. (2012). A comparison of performance of analytic vs holistic scoring rubrics to assess L2 writing. *Iranian Journal of Language Testing*, *2*(1), 59-92. https://www.ijlt.ir/article_ 114361_9544f0e7ef140d3731098f945f34a848.pdf

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

Zorbaz, K. Z. (2013). Yazılı anlatımın puanlanması. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *13*(1) 179-192. https://openaccess.mku.edu.tr/xmlui/bitstream/handle/20.500.12483/1993/Zorbaz%2c%2Kemal%20Zeki%202013.pdf?sequence=1&isAllowed=y

JLERE
JOURNAL OF LANGUAGE EDUCATION AND RESEARCH

Dil Eğitimi ve Araştırmaları Dergisi, 2023, 9 (2), 364-406

Araştırma Makalesi

Yazma Becerisinin Değerlendirilmesinde Puanlayıcı Davranışlarının İncelenmesi

Yusuf Polat* **Nejla Gezmiş****

MAKALE BİLGİSİ

Geliş: 16.07.2023
Düzeltme: 23.08.2023
Kabul: 27.08.2023
Doi: 10.31464/jlere.1328127

Anahtar Sözcükler:
Yabancı dil öğretimi
yazma becerisi
ölçme ve değerlendirme
puanlayıcı davranışları

ÖZET

Çalışmanın amacı yabancı dil öğretmenlerinin, yazma becerisinin değerlendirilmesindeki puanlama davranışlarına ilişkin betimsel veriler elde etmektir. Nicel araştırma yönteminin kullanıldığı çalışmada yarı yapılandırılmış görüşme tekniği ile 73 öğretmenden demografik bilgiler, İngilizce ve Fransızca dillerini öğrenen öğrencilere yazdırılan B1 düzeyindeki bir metne verilen puan ve bu metni değerlendirme sürecini betimledikleri paragraf elde edilmiştir. Verilerin SPSS programı, belge inceleme ve betimsel çözümleme teknikleriyle incelenmesi sonucunda öğretmenlerin puanlama davranışlarının farklılık sergilediği görülmüştür. Nitekim öğretmenlerin puanlama yaparken çoğunlukla ölçüt kullanmadıkları belirlenmiştir. Ayrıca ölçüt kullanan öğretmenlerin, farklı sayılarda bölümlerden oluşan ve farklı veya aynı puan değerine sahip ölçekler geliştirmesi öğretmenlerin kararsızlık yaşadığını göstermesi bakımından önemlidir. Öğretmenlerin çoğunlukla yazma ürününün biçimsel boyutuna odaklandığı, sınırlı bir bölümünün hata türleri arasında ayrım yaptığı veya hata ile puan denkliğini gözettiği sonucuna ulaşılmıştır. Katılımcıların cinsiyeti, çalıştığı kurumu ile verdiği puan arasında bir ilişki olduğu, ancak yaşı ve deneyim süresi ile verdiği puan arasında anlamlı bir ilişki olmadığı sonucuna da ulaşılmıştır.

Bilgilendirme

Yazılı anlatım ürünlerini zaman ayırarak değerlendiren öğretmenlerimize çalışmamızın veri toplama aşamasına sağladıkları katkıdan ve emeklerinden dolayı teşekkür ederiz.

Yayın Etiği Bilgilendirme

Bu çalışma için Etik Kurul onayı alınmıştır: Kırıkkale Üniversitesi Sosyal ve Beşeri Bilimler Araştırmaları Etik Kurulu, 18/10/2022 tarih ve 09 no.

Yazarların Katkı Oranı

Yazarlar eşit oranda katkı sağlamıştır.

Çıkar çatışması

Bu çalışmada çıkar çatışması yoktur.

Gönderim

Polat, Y., & Gezmiş, Y. (2023). Yazma becerisinin değerlendirilmesinde puanlayıcı davranışlarının incelenmesi. *Journal of Language Education and Research, 9*(2), 364-406.

* Prof. Dr., ORCID ID: https://orcid.org/0000-0001- 9341-6643, Kırıkkale Üniversitesi, Fransızca Mütercim Tercümanlık Anabilim Dalı, polatyus@gmail.com

** Dr. Öğrt. Üyesi, ORCID ID: https://orcid.org/0000-0003-4909-1460, Kırıkkale Üniversitesi, İngilizce Mütercim Tercümanlık Anabilim Dalı, nejlagezmis@gmail.com

Giriş

Yazma becerisi bireyin öncelikle iletişim amacıyla yazılı bir söylem üretme sürecidir. İletişim şemasından hareketle tanımlamak gerekirse vericinin belli bir bağlamda, bir alıcıya, ortak bir düzgü (İngilizce, Fransızca, Türkçe vb.) kullanarak oluşturduğu yazılı, diğer bir deyişle görsel iletiyi bir oluktan (kâğıt, ekran vb.) iletmesine dayanır. Yabancı dil öğretim programlarındaki dört temel beceriden biri olarak yazma becerisi diğer yanlarının dışında ölçme ve değerlendirme açısından önemli bir yer tutar.

Yazma becerisinin değerlendirilmesinde biri doğrudan diğeri dolaylı olmak üzere iki ana yaklaşım söz konusudur. İlgili becerinin tüm yönlerinin göz önünde bulundurulduğu bir değerlendirme süreci olan doğrudan yaklaşımda (Coombe, Folse ve Hubley, 2007) öğrenenin yazılı dilde iletişim kurma becerisi değerlendirilmektedir. Bu tür değerlendirmede; klasik ölçme araçları kullanılırken öğrenenin sınıf içerisinde belli bir zamanda bir içerik üretmesi, fikirlerini düzenlemesi, uygun ve doğru söz varlığını, dilbilgisel ve sözdizimsel bilgisini kullanması istenmektedir. Bu değerlendirme türünün sınırlı zamanda her ortamda uygulanabilir olması uygulanmasının kolay olmasını sağlar, fakat öğrenen üzerinde zaman baskısı yaratması, yazma sürecinin hazırlık yapma, plan oluşturma, taslak yazma ve gözden geçirme aşamalarının tam olarak uygulanmasına izin vermemesi gibi olumsuz yönleri de vardır. Dolaylı değerlendirmede, kullanılan dilin iletişim açısından yerine getirdiği işleve değil, dilbilgisel bakımdan doğru olup olmadığına odaklanılmaktadır (Coombe, Folse ve Hubley, 2007). Dolayısıyla, bu tür değerlendirmede çoktan seçmeli testler gibi ölçme araçları kullanılarak tümce düzeyinde noktalama, imla, dilbilgisi gibi bileşenlerin doğru kullanımına yönelik değerlendirme yapma olanağı sunulmaktadır. Değerlendirme sürecinde son yıllarda kullanılmaya başlanan süreç odaklı ölçme araçlarında ise yazma becerisinin ölçülmesi ve değerlendirmesi süreç içerisine yayılmakta ve öğrencinin bu süreçteki gelişimi izlenmektedir. Zaman açısından oldukça zahmetli ve uzun bir sürece yayılan bu ölçme yönteminde daha güvenilir sonuçlar alınmaktadır (Brown, 1989).

O'Malley ve Pierce (1996) gibi araştırmacılar, yazma becerisine yönelik ölçme ve değerlendirme işleminin öğrenenlere çeşitli konularda yazma ödevi verilmesi ve bu ürünün, taşıdığı ileti, açıklık ve yazım, noktalama gibi mekanik boyutları bakımından değerlendirilmesi gerektiğini savunmuşlardır. Diğer bir deyişle, yazma becerisinin değerlendirilmesinde yazma sürecinin karmaşık yapısı göz önüne alınarak tek boyutlu değil, çok boyutlu bir ölçme yapılması önerilmektedir. Ancak, Türkiye'de gerek anadili gerekse yabancı dil öğretimi alanında yazma eğitiminin en önemli sorunlarından biri kullanışlı ve tutarlı bir ölçme değerlendirme çalışmasının yapılamamasıdır (Karatay, 2011). Yazma, dilin sadece söz varlığı, dilbilgisi, yazım ve noktalama bilgisini içermediğinden ölçülmesinde kullanılacak olan yaklaşımların da daha kapsamlı olması ve yazma becerisinin karmaşık süreçlerini içermesi gerekmektedir. Kalay ve Büyükkarcı (2020), Türkiye'deki 97 İngilizce öğretmeniyle anket ve görüşme yoluyla gerçekleştirdikleri çalışmalarında öğretmenlerin çoktan seçmeli test, boşluk doldurma, eşleştirme, doğru yanlış, soru-cevap gibi klasik ölçme araçlarının yanı sıra daha karmaşık süreçleri işaret eden paragraf/kompozisyon yazma, dönem ödevi, öğrenci günlükleri, akran

değerlendirme, öğrenci dosyası değerlendirme, standart yazma testleri gibi süreç odaklı araçları da kullandıklarını belirlemişlerdir.

İster klasik ister süreç odaklı olsun yazma becerisinin etkili bir biçimde değerlendirilmesi ölçme aracının güvenirlik, geçerlilik ve uygulanabilirlik (Weigle, 2007) özelliklerine bağlıdır. Öte yandan, yazma becerisinin değerlendirilmesinde güvenirlik genellikle diğer dil alan ve becerilerine göre daha düşüktür (Calp, 2013). Nitekim Mousavi (2002) de yazma becerisinin değerlendirilmesinde ölçme ve değerlendirme aracının önemi üzerinde durmaktadır. Yazma görevini belirleme ve ona uygun yönerge yazma ölçme ve değerlendirmenin geçerliliğini etkileyen temel unsurların başında gelmektedir. Ölçülmesi hedeflenen yazma becerisinin yönünü tam ve açık olarak ortaya çıkaracak bir görev verilmesi ve bu göreve süreç boyunca öğrenenden beklenen davranışları açık bir şekilde anlatan yönerge eklenmesi özel bir önem taşımaktadır. Yönergeler yalnızca öğrenciler için değil, puanlayıcılar açısından da önemli bir işleve sahiptir. Puanlayıcıların bu yönergeleri ne ölçüde dikkate aldığı, öğrenenin yönergeye uygun yazıp yazmadığına bakılması ve ortaya çıkan ürünün nesnel bir yaklaşımla değerlendirilmesi ölçme ve değerlendirmenin tutarlı ve güvenilir olması açısından önemlidir.

Yazılı anlatım ürünlerinin çok çeşitli oldukları göz önüne alındığında her türün kendine özgü gereklilikleri vardır. Dolayısıyla değerlendirme sürecinde bu gerekliliklerin sağlanıp sağlanmadığının saptanması gerekmektedir. Bu nedenle, puanlama davranışının gerçekleştirilme biçimi, alanyazında yazma becerisini değerlendirmeye yönelik tartışılan konulardan bir diğeridir. Yaygın olarak kullanılan belli başlı üç puanlama yöntemi bulunmaktadır: Çözümleyici, bütüncül ve temel özelliklere göre puanlama (Cooper, 1984; Perkins, 1983; Stiggins ve Bridgeford, 1983; Weigle, 2007; Wiseman, 2012; Zorbaz, 2013). Çözümleyici puanlamada kullanılan ölçek, diğer bir deyişle rübrik, yazma becerisini oluşturan her bir unsurla ilgili ayrı ayrı değerlendirme yapılmasına olanak tanır. Öğrenenin güçlü ve zayıf yönleri konusunda ayrıntılı bilgi sağlarken yazılı anlatım ürününün ayrıntılı bir şekilde incelenmesini ve eksikliklerin işaretlenmesini gerektirir. Bu nedenle puanlayıcılar için zaman alıcı ve uğraştırıcıdır. Buna karşılık bütüncül puanlamada kullanılan ölçek, puanlayıcıların yazılı ürüne dair genel yargılama yapmalarını sağladığı için zaman ve enerji bakımından daha tasarrufludur. Temel özelliklere göre puanlama, genellikle söylem türüne özgü bir değerlendirme sunar. Bu tür ölçekte, söylem ve metin türünün özelliklerine yönelik maddelere yer verilir. Temel özelliklere göre puanlama ölçeğinin avantajı, türe özgü yazılı üretimde bulunulduğunda öğrenenin türün özelliklerine dair bilgisinin ve becerisinin yoklanmasını sağlaması ve öğrenenlerin söylem türüne özgü eksikliklerinin tespitini kolaylaştırmasıdır. Temel özelliklere göre puanlama ölçeğinin hazırlanması daha uğraştırıcı, zaman ve uzmanlık gerektiren bir uğraştır. Bu nedenle diğer iki ölçeğe göre kullanımı daha azdır.

Alanyazında puanlama yöntemlerini konu alan çalışmaların çokluğu ve sonuçlarının farklılığı dikkat çekmektedir. Wiseman (2012) özel olarak puanlama eğitimi almış beş öğretmenin İngilizceyi yabancı dil olarak öğrenen öğrenciler tarafından üretilen 60 yazılı anlatım ürününü çözümleyici ve bütüncül ölçeklerle puanlaması şeklinde gerçekleştirdiği çalışmasında çözümleyici puanlamanın öğrenenler arasındaki farklılıkları daha ayrıntılı ve iyi bir şekilde ortaya çıkardığını göstermiştir. Sakyi (2000) ise

çalışmasında bütüncül puanlama yapan puanlayıcıların kompozisyonun genelini değerlendirirken dil yapılarına ve içeriğe daha fazla odaklandıklarını savunmaktadır. Polat (2003) ve Turgut (1990), çözümleyici puanlamanın, puanlayıcılar arasında uyum sağladığı için daha güvenilir olduğunu iddia ederken Oruç (1999), bütüncül puanlama ile puanlayıcılar arasındaki tutarsızlıkların azaldığını savunmuştur. Öte yandan Han ve Huang (2017), 10 puanlayıcı ile gerçekleştirdikleri çalışmalarında hem puanlar hem de güvenirlik açısından iki puanlama türü arasında anlamlı bir fark olmadığını ve puanlayıcıların her iki türün de avantajları ve dezavantajları olmasına karşın bütüncül puanlama ölçeğini tercih ettiklerini göstermişlerdir. Buna karşılık, alanyazında tam tersi sonuçlar veren çalışmalar da vardır. Örneğin Ghalib ve Hattami (2015), Yemen'de İngilizcenin yabancı dil olarak öğretilmesi bağlamında gerçekleştirdikleri çalışmalarında bütüncül puanlamanın çözümleyici puanlamaya göre daha yüksek olduğunu, puanlayıcılar arasındaki farklılıkların bütüncül puanlamada daha fazla olduğunu göstermiş ve çözümleyici puanlamanın daha güvenilir ve tutarlı ölçme yaptığını iddia etmişlerdir. Aynı şekilde Asassfeh (2021) de yabancı dil öğretimi alanında 48 öğretmenin bütüncül ölçek kullanarak yaptıkları puanlamanın çözümleyici ölçek kullanarak yaptıkları puanlamaya göre daha yüksek olduğunu göstermiş ve puanlayıcıların çözümleyici ölçek kullandıklarında ayrıntılara odaklandıkları için puanlarının düştüğünü iddia etmiştir.

Alanyazında puanlama yöntemleri konusundaki çalışmalar çeşitlilik göstermesine ve farklı sonuçlar ortaya koymasına rağmen ölçek kullanılmasının yazma becerisinin değerlendirilmesi sürecine nesnellik kazandırdığı açıktır. Crusan, Plakans ve Gebril (2016) tarafından 41 ülkede ikinci dilde yazma dersi veren 701 eğitimciyle gerçekleştirilen çalışma; katılımcıların %80'inin yazma becerisini değerlendirirken ölçek kullandığını ve katılımcıların yarıya yakınının ölçek kullanmanın öğrencilerin aldıkları notu anlamlandırmada etkili olduğunu düşündüklerini ortaya koymaktadır. Puanlama ölçeği yalnızca öğretmenler için değil öğrenenler için de önemli bir bileşendir. Öğrenenlere ödev veya ölçme aracı olarak bir yazma görevi verilirken bu görevin değerlendirilmesinde kullanılacak ölçütlerin belirlenmesi ve bunların bir ölçek halinde öğrenene sunulması da büyük önem taşımaktadır. Seviour (2015) değerlendirme ölçeğinin öğrenciler için açık ve ulaşılabilir olması, bu ölçek aracılığıyla öğrencilerden beklenenlerin neler olduğu, üretilen yazılı ürünün nasıl değerlendirileceği ve kabul edilebilir bir yazılı ürünün hangi niteliklere sahip olacağı konusunda bilgilendirilmeleri gerektiğini vurgulamaktadır. Thomas (2020) ise değerlendirme amacıyla bir ölçek geliştirirken veya seçerken, ölçeğin güvenilir ve geçerli olması için bu ölçekle neyin ölçüleceğinin, hangi amaçla ölçme yapılacağının, ne kadar zamanda uygulanabileceğinin ve ölçeği kullanacak kişiler için bir eğitimin gerekli olup olmadığına dikkat edilmesi gerektiğini savunmaktadır.

Puanlayıcı Davranışları

Yazma becerisinin değerlendirilmesinde etkili olan diğer bir etmen ise puanlayıcılardır. Smith (1993) araştırmasında, puanlayıcıların bilgisinin ve kullandığı ölçme tekniğinin ölçmenin güvenirliğini etkilediğini belirtmiştir. Bu şaşırtıcı bir durum değildir, çünkü yazma becerisinde üretilen ürün karmaşık olup bu ürünü puanlayacak olan kişinin mesleki deneyimi, eğitimi, yazılı anlatım ürününün nasıl olması gerektiği

konusundaki görüşleri puanlama sistemi üzerinde etkilidir. Bu nedenle, puanlayıcıların ölçme ve değerlendirme konusunda nitelikli bir eğitim almış olmaları gerekmektedir. Köksal (2004) çalışmasında, Türkiye'de öğretmenlerin yazma becerisinin değerlendirilmesi konusunda yeterince eğitim almadıklarını ve öğretim programında sunulan yollarla yazma becerisini değerlendirme eğilimi sergilediklerini ileri sürmüştür. Nitekim dil öğretimi alanında öğretmen yetiştiren kurumların programlarına bakıldığında becerilerin ayrı ayrı öğretimi konusunda eğitim verilirken becerilere yönelik ölçme ve değerlendirme eğitiminin bulunmadığı, genel anlamda bir ölçme ve değerlendirme eğitiminin verildiği görülmektedir. Uluslararası alanyazında da değerlendirme konusunda bilgi eksikliği bulunduğunu ve öğretmen yetiştirmenin ölçme ve değerlendirme yönünün yetersiz olduğunu iddia eden araştırmalar (Brown ve Bailey, 2008; Mertler, 2009; Popham, 2009; Weigle, 2007) söz konusudur. Dolayısıyla dil öğretmenleri arasında yazma becerisi bakımından ölçme ve değerlendirme konusunda okuryazar olma oranının düşük olduğu belirtilebilir. Nitekim Mede ve Atay (2017) da 350 öğretmen ile gerçekleştirdikleri çalışmada öğretmenlerin bütünleşik beceriler içindeki üretim ve alımlama becerilerine yönelik ölçme ve değerlendirme konusunda kendilerini yetersiz gördüklerini bulgulamıştır.

Ancak güvenilir bir ölçme değerlendirme yapılabilmesi için süreci gerçekleştiren kişilerin, yani puanlayıcıların -okul ortamında öğretmenlerin- ölçme ve değerlendirme konusunda oldukça yetkin olmaları gerekmektedir. Crusan (2010) öğretmenlerin biçimleyici ve özetleyici değerlendirme türleri arasındaki farkları bilmesi, farklı amaçlar için gerekli olan verilerin sunumunu sağlayacak yönergeleri yazma becerisine sahip olması, kullanılan ölçütlerin önceliklerini bilmesi ve değerlendirmenin önemini kavraması gerektiğini vurgulamaktadır. Aynı şekilde Weigle (2007) de yazma becerisinin değerlendirilmesi için öğretmenlerin yazma etkinlikleri düzenleme, yönetme ve puanlama becerilerine sahip olmaları gerektiğinin altını çizmektedir. Sonuç olarak, puanlayıcılardan ölçme ve değerlendirme amacıyla bir araç geliştirebilmeleri, bu aracı uygulayabilmeleri ve puanlayabilmeleri beklenmektedir.

Puanlama sürecindeki davranışlar puanlayıcının kişiliğinin ve eğitiminin yanı sıra Baker (2016)'ın da belirttiği gibi puanlayıcının, öğrencinin çabasını takdir etme isteği ve öğrencinin anlatmak istediğini anlayabilmesi gibi etmenlerden de etkilenebilmektedir. Bununla birlikte, puanlayıcılar, yazılı anlatım ürünlerini geçerli, nesnel, adil, açık, sistematik, ölçüte dayalı ve güvenilir bir ölçme süreci ve yöntemi ile değerlendirmelidir. Puanlayıcıların özellikle yazma becerisinin değerlendirilmesinde bir ölçek kullanması gerektiğinin altı birçok araştırmada çizilirken Lumley (2002) yaptığı araştırmasında puanlayıcıların değerlendirme ölçeğini kullanma şekillerinin oldukça tutarsız olduğunu göstermiştir. Benzer şekilde, istatistiksel olarak Rasch modeli kullanılarak yapılan pek çok çalışmada (Du ve Wright, 1997; Du, Wright ve Brown, 1996; Engelhard, 1994; Lunz, Wright ve Linacre, 1990) farklı puanlayıcıların aynı eğitimi alsalar ve aynı ölçeği kullansalar bile farklı puanlama yaptıkları ortaya çıkmıştır. Puanlayıcıların ölçme ve değerlendirme sürecinde önceliklerinin ve beklentilerinin ve yazılı anlatım ürünü üzerinde odaklandıkları boyutların farklı olması bu durumun altında yatan etmenler arasındadır.

Puanlayıcının karar verme sürecinde yaptığı hatalar ölçme ve değerlendirmenin geçerliliğini ve güvenirliğini etkiler (Erman Aslanoğlu ve Şata, 2021). Puanlayıcıların;

yazılı anlatım ürünün değerlendirmeleri sürecinde cinsiyet, ırk, deneyim, uzmanlık gibi öğrenen grubunun çeşitli özelliklerinden dolayı veya el yazısı, kâğıt düzeni, yazılı üründe savunulan görüş nedeniyle daha düşük/yüksek puan takdir etmeleri değerlendirme sürecindeki puanlayıcı etkisini göstermektedir. Gyagenda ve Engelhard (2009) yaptıkları çalışmada puanlayıcıların, öğrencilerin cinsiyetine göre farklı puanlama davranışı sergilemediğini gösterirken Engelhard ve Myford (2003) puanlayıcıların öğrencilerin cinsiyetine, ırkına ve en başarılı olduğu dile göre farklı puanlama yaptıklarını bulgulamıştır. Johnson ve Lim (2009) yaptıkları incelemede anadili konuşuru olan puanlayıcıların puanları ile anadili konuşuru olmayan puanlayıcıların puanları arasında az da olsa fark olduğunu belirlemiştir. Erman Aslanoğlu ve Şata (2021) anadilinde yazılı anlatım ürününün değerlendirilmesi sürecinde puanlayıcıların, öğrencilerin genel akademik başarı düzeylerini göz önüne aldıklarını, ancak cinsiyetlerini göz önüne almadıklarını ve devlet kurumlarında ya da özel okullarda çalışan öğretmenlerin farklı puanlama davranışları sergilediklerini göstermiştir.

Birçok farklı bileşenin göz önünde bulundurulmasını gerektiren karmaşık işlemler sonucunda ortaya çıkan yazılı anlatım ürününün değerlendirilmesi de karmaşık ve sorunlu bir süreçtir. Bu sorunlardan ilki, işlemin yol açtığı iş yükünün fazlalığıdır. Öğretmenlerin okuldaki iş yüklerinin, sınıftaki öğrenci sayılarının fazla olması ve yazılı anlatım ürünlerinin değerlendirilmesi sürecinin uzun zaman ve çaba gerektirmesi kapsamlı bir değerlendirme yapmayı güçleştirmektedir. İkincisi, üretilmesi beklenen ürün için kesin ve tam bir doğru cevabın bulunmamasına bağlı olarak yazılı anlatımların değerlendirilmesinde tamamen nesnel ölçütler kullanılmasının oldukça zor olmasıdır. Üçüncüsü, puanlayıcıların, kendilerine bir ölçek verilmesine rağmen, ölçekte yer verilen boyutlara olduğu kadar kendi içsel değerlendirme ölçütlerine de odaklanabiliyor olmalarıdır (Li ve Huang, 2022). Oysa yazılı anlatım ürünlerinin de mümkün olduğunca nesnel biçimde değerlendirilmesi gerekmektedir. Uygulama alanına bakıldığında öğretmenlerin yazılı anlatım ürünlerini kendi uzmanlıklarını temel alarak ve kâğıttan edindikleri izlenimlere göre puanladıkları (Çetin, 2002) ileri sürülmektedir. İzlenim ile değerlendirme sürecinde genellikle göze çarpan mekanik hatalar işaretlenmekte, yazının veya kâğıdın düzenine önem verilmektedir. Göçer (2011), Kayseri'de görev yapan 12 Türkçe öğretmeni ile görüşme yaparak gerçekleştirdiği durum araştırmasında, yazma becerisinin değerlendirilmesine ilişkin öğretmen görüşlerini derlemiştir. Çalışmada öğretmenlerin yazma becerisini değerlendirmede bütünleşik ve sürece yayılmış bir değerlendirme yerine toplu bir değerlendirme yaptıkları, çoğunun kompozisyon yazmanın yanı sıra farklı araçlarla değerlendirme yaptığı ve değerlendirme sürecinde zaman ve uygulama biçimi konusunda sıkıntı yaşadıkları belirlenmiştir. Ayrıca yazma becerisinin değerlendirilmesinde ortak bir değerlendirme ölçeği kullanmadıkları, değerlendirme sürecinde plan, yazının düzgünlüğü, kâğıdın düzeni, yazım ve noktalama gibi biçimsel niteliklere dikkat ettikleri bulgulanmıştır.

Özetlemek gerekirse yazma becerisinin değerlendirilmesinde nesnelliği etkileyen etmenler; puanlama, puanlayıcı ve puanlama yöntemlerini de içeren değerlendirme süreci; öğrencinin cinsiyeti, yaşı, ırkı, etnik yapısı, ait olduğu sosyal sınıf, öğrenme ortamları gibi özellikleri; yazma göreviyle ilgili ödevin kendisi, yönergesi, ölçekteki boyutlar gibi

faktörler (Gyagenda ve Engelhard, 2009) olarak karşımıza çıkmaktadır. Bundan başka, yazma becerisinin değerlendirilmesinin sorun teşkil ettiğini gösteren çok sayıda araştırma (Calp, 2013; Cole, Haley ve Muenz, 1997; Hamp-Lyons, 2002; White, 1994) mevcut olup bu sorunlar arasında ölçme ve değerlendirme yöntemi (Beck, Llosa, Black, ve Anderson, 2018; Cooper, 1984; Han ve Huang, 2017; Şeker, 2018; Tokur Üner ve Aşılıoğlu, 2022; Wilson vd., 2016), ölçme ve değerlendirme araçlarının geçerliği ve güvenirliği (Brown, Glasswell ve Harland, 2004; O'Neill, 2011) ve ölçme ve değerlendirmeyi gerçekleştiren kişilerin tutarlılığı ve güvenirliği (Erman Aslanoğlu ve Şata, 2021; Gyagenda ve Engelhard, 2009; Lumley, 2002; Smith, 1993; Wind ve Engelhard, 2012; Zhang, 2016) sayılmaktadır.

Alanyazında yazma becerisinin değerlendirilmesi konusunda yapılan araştırmaların büyük çoğunluğu değerlendirme aracının özellikleri, değerlendirme yöntemi ve değerlendirme sırasında kullanılan ölçek türleri üzerinde yoğunlaşmaktadır. Araştırmalarda katılımcılara değerlendirme veya puanlama esnasında kullanmaları için bir ölçek verilerek katılımcıların yönlendirildiği görülmektedir. Ayrıca, Türkiye'de okutulan ders kitaplarının da ölçekler barındırdığı bilinmektedir. Araştırmalarda ayrıca yönlendirmeler ışığında öğretmenlerin olağan değerlendirme süreçlerinde de ölçek kullandığı varsayılmaktadır. Alanyazında öğretmenlerin ölçme ve değerlendirme yöntem ve tekniklerini araştıran çok sayıda araştırma olmasına rağmen öğretmenlerin puanlama davranışlarını betimleyen çalışmaların sayısı sınırlıdır. Bunun yanı sıra, öğretmenlerin yönlendirilmediğinde nasıl bir yol izleyerek değerlendirme yapacakları konusunda bilgi içeren araştırmaya da rastlanmamıştır. Bu nedenle bu araştırmada yabancı dil öğretmenlerinin yazma becerisinin değerlendirilmesinde sergiledikleri puanlama davranışlarının incelenmesi amaçlanmıştır. Bu amaç doğrultusunda araştırma kapsamında aşağıda sunulan iki soru ve alt sorulara yanıt aranmaktadır:

1) Yazma becerisinin değerlendirilmesi kapsamında yabancı dil öğretmenleri, üretilen bir metni puanlarken hangi puanlayıcı davranışlarını sergilemektedir?

a) Yazma becerisinin değerlendirilmesi kapsamında yabancı dil öğretmenlerinin, üretilen bir metne verdikleri puanlar ile öğretmenlerin demografik özellikleri arasında istatistiksel olarak anlamlı bir ilişki var mıdır?

b) Yazma becerisinin değerlendirilmesi kapsamında yabancı dil öğretmenleri, üretilen bir metni puanlarken hangi tür işaretleme davranışlarını sergilemektedir?

c) Yazma becerisinin değerlendirilmesi kapsamında yabancı dil öğretmenleri, üretilen bir metni puanlarken öğrenci hataları konusunda nasıl bir yol izlemektedir?

d) Yazma becerisinin değerlendirilmesi kapsamında yabancı dil öğretmenleri, üretilen bir metni puanlarken ürünün hangi boyutlarına odaklanmaktadır?

2) Yazma becerisinin değerlendirilmesi kapsamında yabancı dil öğretmenleri, üretilen bir metni puanlama süreçlerini nasıl betimlemektedir?

Yöntem

Yabancı dil öğretmenlerinin yazma becerisinin değerlendirilmesinde puanlama yaparken sergiledikleri davranışlarının betimlenmesinin amaçlandığı bu çalışmada öğretmenlerden, mütercim tercümanlık öğrenimi görmekte olan öğrencilere yazdırılan ve rastgele örnekleme yöntemiyle seçilen İngilizce ve Fransızca B1 düzeyindeki aynı paragrafı puanlamaları ve puanlama süreçlerini betimlemeleri istenmiştir. Öğretmenlerin puanlama yaptıkları kâğıtlar incelenerek ve çeşitli değişkenlere göre karşılaştırılarak puanlama davranışlarının belirlenmesi amaçlanmıştır. Ayrıca öğretmenlerin kendi değerlendirme süreçlerini betimledikleri paragrafların incelenmesi ile de yazma becerisinin değerlendirilmesinde nasıl bir yol izlediklerinin anlaşılması hedeflenmiştir.

Araştırmanın Katılımcıları

Araştırmaya puanlayıcı olarak toplam 73 öğretmen gönüllü olarak katılmıştır. Çoğunluğunu kadın öğretmenlerin oluşturduğu katılımcılar genellikle 31-40 yaş aralığındadır. Yabancı dil öğretmeni olan katılımcıların çoğunluğu İngilizce öğretmenliği bölümü mezunudur. Ayrıca öğretmenlerin büyük bir çoğunluğu formasyon eğitimi almıştır ve 15 yıl üstü bir deneyime sahiptir. Ek olarak, çalıştıkları düzey genellikle üniversite olup devlet okullarında görev yapmaktadırlar. Katılımcıların demografik özelliklerine ilişkin ayrıntılı veriler Çizelge 1'de sunulmuştur.

Çizelge 1. Katılımcılara İlişkin Bilgiler

Değişken	Kategori	s	Yüzde (%)
Cinsiyet	Erkek	15	20.5
	Kadın	58	79.5
Yaş	20-30 arası	11	15.1
	31-40 arası	30	41.1
	41-50 arası	24	32.9
	51 ve üstü	8	11.0
Mezuniyet	İngilizce Öğretmenliği	48	65.8
	Diğer	25	34.2
Formasyon	Var	71	97.3
	Yok	2	2.7
Deneyim	1-5 yıl	6	8.2
	6-10 yıl	15	20.5
	11-15 yıl	14	19.2
	16-20 yıl	17	23.3
	21 yıl ve üstü	21	28.8
Kademe	İlkokul	6	8.2
	Ortaokul	7	9.6
	Lise	20	27.4
	Üniversite	35	47.9
	Birden çok	5	6.8
Kurum	Devlet	62	84.9
	Özel	11	15.1

Veri Toplama Araçları

Araştırmanın verileri yarı yapılandırılmış görüşme tekniği ile elde edilmiştir. Bu amaçla hazırlanan veri toplama formu üç bölümden oluşmaktadır. Birinci bölümde katılımcı öğretmenlerin cinsiyet, yaş, mezun oldukları program, mesleki deneyim, şu anda çalışılan öğretim düzeyi, branşı ve kurum bilgileri gibi demografik özelliklerine yönelik sorulara yer verilmiştir. İkinci bölümde öğretmenlerden verilen paragrafı 100 üzerinden puanlamaları, üçüncü bölümde ise puanlama yapma süreçlerini yazılı olarak kısaca anlatmaları istenmiştir.

Yayın Etiği

Görüşme formuna yönelik etik kurul izni Kırıkkale Üniversitesi Sosyal ve Beşerî Bilimler Araştırmaları Etik Kurulu'ndan 18/10/2022 tarih 09 no'lu karar ile alınmıştır.

Verilerin Çözümlenmesi

Elde edilen verilerin çözümlenmesinde belge inceleme ve betimsel çözümleme tekniği kullanılmıştır. Çözümlemenin ilk aşamasında çalışma grubunu oluşturan öğretmenlerin demografik özelliklerine odaklanılmış ve veri toplama aracı olarak kullanılan öğretmen formunun ilk bölümünden belge inceleme tekniği ile elde edilen veriler yorumlanmıştır. İkinci aşamada, veri toplama formunun ikinci bölümüyle ilgili gözlemler gerçekleştirilmiş ve belge inceleme ve betimsel çözümleme yoluyla katılımcı öğretmenlerin, paragrafı puanlama şekilleri araştırmacılar tarafından incelenmiş ve öğretmenlerin kâğıt üzerinde gösterdikleri davranışlar belirlenmiştir. Bu davranışlar öncelikle kodlanmış ve sonrasında araştırma sorularından yola çıkılarak oluşturulan tematik çerçeve içerisinde ilgili temaya yerleştirilmiştir. Veri inceleme sürecinin üçüncü aşamasında ise öğretmen formunun üçüncü bölümünden yararlanılmıştır. Bu amaçla öğretmenlerin kendi değerlendirme süreçlerini betimledikleri paragraflar araştırmacılar tarafından okunmuş ve ifade ettikleri puanlama davranışları belirlenmiştir. Aynı şekilde, ifade edilen bu davranışlar kodlanarak yukarıda bahsedilen tematik çerçeveye yerleştirilmiştir. Dolayısıyla, araştırma kapsamında öğretmenlerin demografik özellikleri, puanlama davranışları ve puanlama davranışlarına yönelik söylemleri olmak üzere üç veri kümesi elde edilmiştir. İkinci ve üçüncü veri kümelerinin karşılaştırmalı incelemeleri birinci veri kümesine dayanılarak yapılmaktadır. Ayrıca ikinci ve üçüncü veri kümesinin incelenmesi ile öğretmenlerin söylemleri ile eylemleri arasında tutarlılık olup olmadığının yanı sıra yaş, cinsiyet, deneyim ve puanlama davranışı arasında ilişki olup olmadığının saptanması amaçlanmıştır. Elde edilen bulgular sayı ve yüzde olarak ve çizelgeler yardımıyla sunulmuştur.

Söz konusu çalışmada yazma becerisinin ölçülmesi için sadece kompozisyon yazma tekniğinin kullanılması araştırmanın en önemli sınırlılığıdır. Kompozisyon yazmada öğrencinin dilbilgisel bilgisinin yanı sıra sözdizimsel, anlambilimsel, kullanımbilimsel ve metinbilimsel bilgisinin de kullanılması gerektiği için her türlü yazma becerisinin bu şekilde en kısa yoldan ölçülebileceği düşünülmüştür. Çalışma kapsamında katılımcı öğretmenlerin puanlama davranışlarının incelenmesi amaçlandığı için farklı öğrenciler tarafından üretilen paragraflardan ziyade İngilizce ve Fransızca olmak üzere birer öğrenci

kompozisyonu kullanılmıştır. Böylece öğrenci değişikliği faktörü çalışma dışında tutulmuştur. Türkiye'de yabancı dil olarak farklı dillerin öğretimi söz konusu olsa da en yaygın öğretilen yabancı dil İngilizce ve Fransızca olduğu için ve uygulanabilirlik ilkesi gereği çalışma bu iki dilin öğretmenleriyle sınırlandırılmıştır.

Bulgular

Puanlayıcıların Davranışlarına Yönelik Bulgular

Katılımcıların verdikleri puanlar incelendiğinde en düşük puanın 40, en yüksek puanın ise 100 tam puan olduğu görülmektedir. Verilen puana ilişkin temel bulgu katılımcı öğretmenlerin puan tercihlerinde önemli farklılık olduğu yönündedir. Puanlara ilişkin sayılar ve oranlar Çizelge 2'de sunulmaktadır.

Çizelge 2. Katılımcıların Verdikleri Puanlar

Puan	s	Yüzde (%)
81-100 arası	41	56.2
61-80 arası	20	27.4
40-60 arası	7	9.6
En az 90	1	1.4
Tam puan	1	1.4
Puan yok	2	2.7
Toplam	73	100.0

Puanlama üzerinde yapılan incelemeler sırasında, 73 öğretmenden ikisi herhangi bir puan vermediğinden çözümlemeye dâhil edilmemiştir. Dolayısıyla 71 öğretmenin vermiş olduğu puanlar üzerinde inceleme yapılmıştır. Buna göre ilk olarak puanların normal dağılım sergileyip sergilemediğine bakılmıştır. Çizelge 3'te görüldüğü üzere, çarpıklık ve basıklık katsayıları -1<p>1 arasında olmadığı için ve Kolmogorov testi sonucu p<0.05 olduğu için veriler normal dağılım sergilememiştir.

Çizelge 3. Puanlamalara İlişkin Verilerin Çarpıklık ve Basıklık Katsayıları

Katılımcı Sayısı	Ortalama	Çarpıklık Katsayısı	Basıklık Katsayısı	Kolmogorov Smirnov Katsayısı
71	82.89	-1.336	1.767	.000

Puanlama verileri normal dağılım sergilemediği için öğretmenlerin cinsiyetlerine göre verdikleri puanlar arasında fark olup olmadığını incelemek üzere veriler Mann-Whitney testi kullanılarak incelenmiştir. Yapılan inceleme sonucunda p<0.05 olduğu için kadın ve erkek öğretmenlerin verdiği puanlar arasında kadınlar lehine anlamlı bir fark olduğu görülmüştür (bkz. Çizelge 4). Sıra ortalamaları ve toplamları daha yüksek olduğu için kadın öğretmenlerin erkeklere göre daha yüksek puan verdikleri anlaşılmaktadır.

Çizelge 4. Puanlamalara İlişkin Verilerin Cinsiyet Değişkenine Göre İncelenmesi

Kategori	Sayı	Ortalama	Sıra Ortalaması	Sıra Toplamı	U	P
Kadın	57	84.67	38.89	2216.50	234.500	.017
Erkek	14	75.64	24.25	339.50		

Çalışma grubunu oluşturan öğretmenlerin devlet veya özel kurumlarda çalıştığı görülmüştür. Çalıştıkları kurum açısından verdikleri puanlar arasında fark olup olmadığını incelemek için yine veriler normal dağılım sergilemediği için Mann-Whitney testi kullanılmıştır. Çizelge 5'te görüldüğü üzere, $p=0.004<0.05$ olduğu için öğretmenlerin verdiği puanlar arasında çalıştıkları kurum değişkenine göre anlamlı bir fark vardır. Sıra ortalamaları ve toplamları daha yüksek olduğu için devlet kurumlarında çalışan öğretmenlerin özel kurumlarda çalışanlara göre puanlamalarının daha yüksek olduğu anlaşılmaktadır.

Çizelge 5. Puanlamalara İlişkin Verilerin Çalışılan Kurum Değişkenine Göre İncelenmesi

Kategori	Sayı	Ortalama	Sıra Ortalaması	Sıra Toplamı	U	P
Devlet	61	85.23	38.87	2371.00	130.000	.004
Özel	10	68.60	18.50	185.00		

Öğretmenlerin puanlamalarının mesleki deneyim sürelerine göre değişiklik gösterip göstermediğinin incelenmesi amacıyla veriler, deneyim süreleri beş kategoriye ayrıldığı için Kruskall Walls testi kullanılarak incelenmiştir. Yapılan inceleme sonucunda $p=0.104>0.05$ olduğu için öğretmenlerin mesleki deneyim sürelerine göre verdiği puanlar arasında anlamlı bir fark olmadığı görülmüştür (bkz. Çizelge 6).

Çizelge 6. Puanlamalara İlişkin Verilerin Mesleki Deneyim Süresi Değişkenine Göre İncelenmesi

Kategori	Sayı	Ortalama	Sıra Ortalaması	sd	Ortalama	P
1-5 yıl	6	73.17	29.25			
6-10 yıl	14	77.43	24.29			
11-15 yıl	13	84.77	39.00	4	7.688	.104
16-20 yıl	17	87.00	42.91			
21 ve üstü	21	84.81	38.29			

Öğretmenlerin puanlamalarının çalıştıkları kademeye göre değişiklik gösterip göstermediğinin incelenmesi amacıyla veriler, kademeler beş kategoriye ayrıldığı için Kruskall Walls testi kullanılarak incelenmiştir. Çizelge 7'de görüldüğü üzere, öğretmenlerin kademelerine göre puanlamaları arasında anlamlı bir fark ($p=0.036<0.05$) vardır. Ortalamalarına bakıldığında birden çok kademede çalışan öğretmenlerin ve üniversite düzeyinde çalışan öğretmenlerin diğerlerinden daha düşük puan verdikleri anlaşılmaktadır.

Çizelge 7. Puanlamalara İlişkin Verilerin Çalışılan Kademe Değişkenine Göre İncelenmesi

Kategori	Sayı	Ortalama	Sıra Ortalaması	sd	Ortalama	P
İlkokul	6	86.33	37.50			
Ortaokul	6	85.83	37.42			
Lise	19	86.11	40.32	4	10.265	.036
Üniversite	35	83.97	37.14			
Birden çok	5	55.40	8.10			

Puanların dışında, bir numaralı araştırma sorusunun b, c, d alt sorularına yanıt vermek üzere katılımcıların değerlendirme yaptıkları kâğıt üzerinde işaretleme davranışları, hata türüne yaklaşımları, öğrenciyle etkileşim tercihleri ve odaklandıkları hatalar gözlemlenmiştir. Bu amaçla a) işaretleme yapılıp yapılmadığı b) yapıldı ise kaç işaret konduğu c) hata türünün belirtilip belirtilmediği d) öğrenciye uyarıda bulunulup bulunulmadığı e) hataların düzeltilip düzeltilmediği f) işaretlemede kodlama yapılıp yapılmadığı g) dilsel bileşen hatalarına dikkat edilip edilmediği (dilbilgisi, noktalama, yazım, sözvarlığı, sözdizim) h) söylem türünün göz önünde bulundurulup bulundurulmadığı ı) içeriğin doğruluğuna dikkat edilip edilmediği i) ölçek kullanılıp kullanılmadığı sorularının yanıtları aranmıştır. Bu yönlere ilişkin gözlem sonuçları Çizelge 8'de ayrıntılı bir şekilde sunulmaktadır. Yapılan inceleme sonucunda, öğretmenlerin genel olarak kâğıt üzerinde hataları düzelttikleri ve dilbilgisi ve ifade yanlışlarına odaklandıkları görülmüştür.

Çizelge 8. Puanlayıcıların Davranışlarına İlişkin Genel Gözlemler

Davranış Kategorisi	s	Yüzde (%)
İşaretleme ile yetinilmiş	30	41.1
Hata türü belirtilmiş	8	11.1
Öğrenciye uyarıda bulunulmuş	11	15.1
Hata düzeltilmiş	47	64.4
İşaretlemede kodlama yapılmış	2	2.7
Dilbilgisi hataları dikkate alınmış	64	87.7
Yazım yanlışları dikkate alınmış	35	47.9
Noktalama yanlışları dikkate alınmış	26	35.6
Sözcük yanlışları dikkate alınmış	31	42.5
İfade yanlışları dikkate alınmış	32	43.8
Söylem türü (mektup) özellikleri dikkate alınmış	12	16.4
İçeriğin doğruluğu göz önünde bulundurulmuş	5	6.8
Ölçek kullanılmış	10	13.7

Puanlayıcıların Söylemlerine Yönelik Bulgular

İki numaralı araştırma sorusunun yanıtlamak üzere katılımcıların puanlama sırasında izledikleri yolu açıkladıkları metinler çözümlendiğinde biri ölçek kullanımı, diğeri ise odaklanılan yön olmak üzere iki ana boyut ortaya çıkmaktadır. Ölçek kullanımına ilişkin boyutta katılımcının ölçek kullandığına ilişkin bilgi verip vermemesi, buna bağlı olarak herhangi bir ölçek çizip çizmemesi, ölçekte puanlama yaparken her bir hataya kaç puan

vereceği ve ölçekteki bölüm sayısına ilişkin veriler elde edilen veriler Çizelge 9'da ayrıntılı olarak sunulmuştur. İnceleme sonucunda öğretmenlerin çoğunluğunun, ölçek çizerek değerlendirmeyi bu ölçeğe göre gerçekleştirdiği görülmüştür. Bununla birlikte, katılımcılar ölçek kullanımına ilişkin farklı söylemlerde de bulunmuştur. Bu bağlamda en göze çarpan fark, Çizelge 9'da da görüldüğü gibi, geliştirilen ölçekteki bölüm sayılarının farklılık göstermesidir. Diğer bir fark ise puanlayıcıların çoğunluğunun ölçekteki bölümlere aynı veya farklı puanları öngörmesidir.

Çizelge 9. Puanlayıcıların Ölçek Kullanımına İlişkin Söylemleri

Ölçek Tercihleri		s	Yüzde (%)
Ölçek kullandığını belirtme		4	5.5
Ölçek çizme		26	35.6
Ölçekteki bölüm sayısı	2	1	1.4
	3	3	4.1
	4	9	12.3
	5	7	9.6
	6	1	1.4
Ölçekteki her bölüme aynı puanı verme		16	21.9
Ölçekteki her bölüme farklı puan verme		10	13.7
Ölçekte puan kullanmama		2	2.7
Ölçekte hata puan ilişkisini söyleme		3	4.1

Katılımcıların söylemlerinden hareketle odaklandıkları veya göz önünde bulundurdukları yönlere ilişkin çözümlemeler çok farklı tercihlerin var olduğunu ortaya koymaktadır. Nitekim puanlayıcılar inceledikleri kâğıtta bütünlük, bağdaşıklık, bağlaşıklık, konuya uygunluk, içerik, dil kullanımı, tutarlılık, anlaşılırlık, dilbilgisi, noktalama, yazım yanlışı, sözcük bilgisi, sözcük sayısı, zaman, hedef kitle, amaç, tür özellikleri, metin bölümleri, düzen, kanıtlar/örnekler, kâğıt düzeni, yazı şekli, yaratıcılık, biçem, akıcılık, dikkat çekicilik, düşünme biçimi, sınıf içi katılım, yaş, planlama ve düzey gibi noktaları göz önünde bulundurduklarını belirtmektedirler. Bunların katılımcı grubu içerisindeki sıklık düzeyleri ile oranlarına bakıldığında yalnızca 1'er katılımcının hedef kitle, sınıf içi katılım, yaş, uyruk ve yaratıcılık sözcükleriyle kodladığımız değişkenleri dikkate aldığı, bu sayının da her bir değişken için %1.4'lük bir orana karşılık geldiği görülmektedir. Hedef kitle üretilen metnin kime yazıldığını, sınıf içi katılım öğrencinin derste gösterdiği performansı belirtirken yaş, metni üretenin çocuk veya yetişkin olup olmadığının, uyruk ise metin üreticisinin öğrenilen dil ile yakınlık ve uzaklık ilişkisinin göz önünde bulundurulduğunu göstermektedir. Katılımcıların göz önünde bulundurdukları değişkenler arasında kâğıt düzeni, akıcılık, düşünme biçimi, konuya uygunluk, bütünlük, kanıtlar/örnekler, tür özellikleri, metin bölümleri, düzey, dil kullanımı, bağdaşıklık, yazım yanlışı, içerik ve noktalama dikkat çekerken metnin düzeni, anlaşılırlık, sözcük bilgisi ve dilbilgisi en fazla üzerinde durulan değişkenlerdir. Sözü edilen değişkenler konusunda katılımcı tercihlerini gösteren sayılar ile oranlar aşağıdaki çizelgede topluca sunulmaktadır.

Çizelge 10. Puanlayıcıların Puanlamada Dikkate Alınan Yön Konusundaki Söylemleri

Odaklanılan Yön	s	Yüzde (%)
Hedef kitle	1	1.4
Sınıf içi katılım	1	1.4
Yaş	1	1.4
Uyruk	1	1.4
Yaratıcılık	1	1.4
Biçem	2	2.7
Tutarlılık	2	2.7
Zaman	2	2.7
Dikkat çekicilik	2	2.7
Planlama	2	2.7
Yazı biçimi	3	4.1
Bağlaşıklık	4	5.5
Sözcük sayısı	4	5.5
Amaç	4	5.5
Kâğıt düzeni	8	11.0
Akıcılık	8	11.0
Düşünme biçimi	8	11.0
Konuya uygunluk	9	12.3
Bütünlük	10	13.7
Kanıtlar/Örnekler	11	15.1
Tür özellikleri	12	16.4
Metin bölümleri	14	19.2
Düzey	14	19.2
Dil kullanımı	15	20.5
Bağdaşıklık	16	21.9
Yazım yanlışı	16	21.9
İçerik	20	27.4
Noktalama	20	27.4
Düzen	22	31.5
Anlaşılırlık	24	32.9
Sözcük bilgisi	35	47.9
Dilbilgisi	68	93.2

Özetlemek gerekirse, elde edilen veriler, araştırma sorularına ilişkin çok sayıda bulguyu ortaya koymaktadır. Birinci bulgu kümesi, öğretmenlerin değerlendirdikleri kâğıt üzerinde yaptıkları işlemleri ilgilendirmektedir. Buna göre, öğretmenler arasında işaretleme yaklaşımları konusunda ortak bir tercih gözlenmemektedir. Kimileri altını çizmek veya daire içine almak gibi işaretleme yaklaşımlarını tercih ederken az sayıda da olsa kimileri hata türünü belirterek işaretlemeyi tercih etmiştir.

İkinci bulgu kümesi öğretmenlerin verdikleri puanları ilgilendirmektedir. Buna göre, katılımcı öğretmenler aynı kâğıda önemli ölçüde farklı puanlar vermektedir. Nitekim aynı kâğıda 40 ile 100 puan arasında değişen puanlar verilmiştir. Üçüncü bulgu kümesi ölçek kullanımına ilişkin uygulamaları ilgilendirmektedir. Buna göre, hiç ölçek kullanmayanlar

olduğu kadar, ölçek kullananlar da vardır. Dördüncü ve son bulgu kümesi cinsiyet, çalışılan kurumun niteliği, deneyim ve çalışılan kademeye göre, verilen puanda fark olup olmadığına ilişkindir.

Son bulgu kümesi katılımcıların ölçme değerlendirmede odaklandığı yöne ilişkin olup bu noktada katılımcıların önemli bir bölümü yazılı anlatım ürününün dilbilgisel boyutuna odaklandığı anlaşılmaktadır. Onu sırasıyla sözcük bilgisi, anlaşılırlık, düzen, noktalama ve içerik boyutları izlerken bunlar da dilbilgisel boyuta dâhildir. Bağdaşıklık, dil kullanımı, metnin bölümleri gibi dilbilgisi dışı boyutlar görece daha az sayıda katılımcı tarafından göz önünde bulundurulmaktadır.

Tartışma

Katılımcı öğretmenlerin puanlama tercihlerinde önemli farklılıklar olduğuna yönelik bulgu birkaç yönden üzerinden durulması gereken bir bulgudur. Öncelikle, katılımcı öğretmenlere herhangi bir ölçek verilmeksizin puanlama yapmalarının beklenmiş olması nedeniyle puanlamada farklılık olması doğal bir sonuç olarak görülebilir. Nitekim alanyazında ölçek verildiğinde bile puanlayıcıların puanlamalarının farklı olduğunu gösteren çok sayıda çalışma (Bachman, 2004; Engelhard ve Myford, 2003; Hunter ve Docherty, 2011; Liu, 2022; Şeker, 2018,) mevcuttur. Öte yandan, bu farklılıkların 40 ile 100 gibi geniş bir aralıkta gerçekleşmesi puanlayıcıların verdikleri puanları güvenirlik bakımından kuşkulu hale getirmektedir. Puanlayıcıların puanlamalarının düşük güvenirlik sergilemesi Gyagenda ve Engelhard (2009) tarafından gerçekleştirilen, 20 puanlayıcı öğrencinin 366 kompozisyonu ölçek kullanarak puanladıkları araştırmanın sonuçlarıyla uyumludur. Çünkü söz konusu araştırmada puanlayıcı eğitimi verilen 20 puanlayıcının verdikleri puanlar arasında güvenirlik katsayısı düşük çıkmıştır. Aynı araştırmadan elde edilen bulgular ile bizim bulgularımız, ölçek verilsin verilmesin veya eğitim almış olsun olmasın öğretmenlerin puanlamada farklı tercihlerde bulunduklarını gösteriyor olmaları bakımından dikkate değerdir. Puanlamadaki bu farklılık oluşturulan ölçekteki boyutların tamamında gözlenirken bizim araştırmamızda da katılımcı öğretmenlerin ölçek oluşturduklarında ölçekteki her bir boyuta aynı veya farklı puan öngördükleri gözlemlenmiştir. Ölçekteki boyutların farklı puanlanması, bir yanıyla puanlayıcıdan diğer yanıyla değerlendirilen boyutun mutlak değer olarak ifade edilemiyor oluşundan kaynaklanabilmektedir. Puanlayıcının, ilgili boyutu önemli görüp görmemesi, değerlendirirken gerekli özeni ve dikkati gösterip göstermemesi puanlayıcıdan kaynaklı farklılıklara yol açabilirken değerlendirilen boyutun doğası gereğince tek ve ülküsel bir yanıtının olmaması da değerlendirilen boyuta bağlı farklılıklara yol açabilmektedir. Öte yandan dilbilgisi gibi kimi boyutlar görece daha mutlak değer olarak ifade edilebilmesine rağmen farklı puanlanıyor olması tüm çabalara rağmen değerlendirme sürecinde puanlama farklılıklarının ortadan kaldırılamayacağını düşündürmektedir. Şeker (2018)'in Türkiye'de bir okulda çalışan üç İngilizce öğretmeni ile gerçekleştirdiği çalışmasında yazma becerisinin değerlendirilmesinde puanlayıcı/öğretmenin davranışları konusunda benzer sonuçlar ortaya koymaktadır. Söz konusu çalışmanın gerçekleştirildiği okulda yazma becerisi önceden hazırlanmış standart bir ölçek ile değerlendirilmiştir. Öğrenciler tarafından üretilen ve okul sistemi içinde sınav olarak kullanılmış olan paragraflardan düşük düzey, orta düzey ve iyi düzeydekiler eşit sayıda olacak biçimde toplamda 75

paragraf seçilmiş ve üç öğretmenden üç gün içinde bu paragrafların 25 tanesini aynı ölçeği kullanarak puanlamaları istenmiştir. Daha sonra diğer 25'ini üç öğretmen tartışarak beraber puanlamıştır ve bu tartışma kayıt altına alınmıştır. Üç hafta sonra ise kalan 25 paragrafı bireysel olarak aynı ölçekle puanlamaları istenmiştir. İlk puanlamadan elde edilen veriler incelendiğinde üç öğretmenin de aynı ölçeği kullanmasına rağmen farklı puanlar verdikleri görülmüştür. Öğretmenlerin ölçekte yer alan dilbilgisel doğruluk, sözlüksel doğruluk, sözdizimsel doğruluk, düzen, mekanik özellikler gibi boyutlarda farklı yargılarda bulundukları anlaşılmıştır. Yapılan istatistiki işlem sonucunda üç öğretmenin ölçekteki maddelere verdikleri puanların birbirleriyle uyumlu olmadığı görülmüştür. Öğretmenlerin beraber tartışarak puanlama yaptıkları oturum kayıtları izlendiğinde öğretmenlerin ilk gün puanlama kararlarının nedenleri konusunda çekince yaşadıkları görülmüştür. İkinci gün öğretmenlerin rahatladığı, birbirleriyle fikir alışverişinde oldukları ve belli boyutlarda ayrı ayrı uzmanlık sergiledikleri tespit edilmiştir. Örneğin dilbilgisel doğruluk konusundaki tartışmalarda bir öğretmenin iddiası dayanak olarak alınırken düzen konusunda başka bir öğretmenin görüşünün temel alındığı bulgulanmıştır. Ayrıca puanlama sürecinin ilk gün daha uzun sürdüğü ancak giderek kısaldığı da araştırma bulguları arasındadır. Süreç sonunda öğretmenlerin yaptıkları puanlamayı daha adil buldukları, tek başına karar vermedikleri için sorumluluğu paylaştıkları anlaşılmıştır. Üç hafta sonra bireysel olarak yapılan puanlama verileri incelendiğinde öğretmenlerin dilbilgisel doğruluk, sözlüksel doğruluk, düzen ve mekanik boyutta benzer puanlama yaptıkları görülmüştür. Çalışma; aynı standart ölçeği kullansalar bile öğretmenlerin aynı kâğıtları farklı şekillerde puanladıklarını; öğretmenlerin bir kısmı yapıya odaklanırken, bir kısmının doğruluğa, bir kısmının da akıcılığa odaklandığını göstermiştir. Şeker (2018)'in çalışmasının dikkat çeken bulgularından biri öğretmenlerin toplamda üç gün süren tartışma sürecinin farklı aşamalarında farklı tepkiler ortaya koymuş olmalarıdır. Nitekim katılımcılar tartışma yaparak gerçekleştirdikleri puanlamanın ilk gününde sessiz ve tereddüt içeren davranışlar sergilerlerken ilerleyen günlerde puanlama konusunda tereddüt hissetmeden fikir tartışmaları yapar ve daha özgüvenli kararlar alır hale gelmişlerdir. Ayrıca öğretmenlerin kullandıkları ölçeği de tartıştıkları ve ölçeğin yetersiz kaldığı yönleri belirledikleri görülmüştür. Daha sonra tekrar yapılan bireysel değerlendirmede öğretmenler arasındaki farklılık azalmış, tartışma yoluyla puanlama kısmında edindikleri deneyim ve bilgileri kendi puanlamalarında kullandıkları görülmüştür. Dolayısıyla çalışma puanlama sürecinde paydaşlarla iş birliği ve tartışma içinde olmanın puanlayıcılar açısından önemli katkıları olduğunu ve değerlendirme sürecine tutarlılık kattığını göstermektedir. Şeker (2018)'in bulgularıyla karşılaştırıldığında araştırmamızdan elde edilen bulgular, araştırmamıza veri sağlayan katılımcıların bireysel olarak önerdikleri ölçeklerdeki farklılıklar ile kimi yönleriyle tutarsız ölçek hazırlama girişimlerinin benzer olduğunu ortaya koymaktadır.

Puanlama tercihlerinde kadın ve erkek puanlayıcılar arasında kadınlar lehine anlamlı bir farklılık bulunması alanyazındaki kimi araştırma verileriyle (Peterson, Childs ve Kennedy, 2004) uyuşmazken cinsiyete bağlı olarak puanlamada farklılıkların bulunduğunun gözlendiği araştırmalar (Gyagenda ve Engelhard, 2009) da mevcuttur. Nitekim Peterson, Childs ve Kennedy (2004) tarafından Kanada'da anadili öğretimi

alanında çalışan 108 öğretmenle gerçekleştirilen araştırmada iki kız ve iki erkek öğrenci tarafından üretilen öyküleyici ve tartışmacı kompozisyonları puanlamaları istenmiş, öğretmenlerin verdiği puanlar arasında puanlayıcının cinsiyetine göre sadece bir kompozisyon fark olduğu, cinsiyetler arası farklılığa dair tutarlı sonuçlar olmadığı ve puanlanan ürünün üreticisinin cinsiyetine bağlı olarak anlamlı bir fark olmadığı gözlenmemiştir. Buna karşılık, Gyagenda ve Engelhard (2009) tarafından yapılan çalışmada öğrencilerin cinsiyeti açısından bakıldığında erkek ve kız öğrenciler arasındaki farklılıkların kızlar lehine anlamlı olduğu görülmüştür. Araştırmada bu farklılığın gerekçesinin, kız öğrencilerin yazma becerisinde daha başarılı olduklarına dair hâkim olan bir inanış veya öğretmenlerin erkek öğrencilerin yazma becerisini geliştirmek için onlar tarafından üretilen ürüne daha çok odaklanması olabileceği belirtilmiştir. Ancak bu araştırmada 20 eğitimli puanlayıcının cinsiyeti incelemelerde dikkate alınmamış ve puanlayıcıların cinsiyet açısından dağılımları verilmemiştir.

Araştırma bulgularımız arasında ölçek oluşturulurken bölüm sayısının ve bölüm başına düşen puanların kimi katılımcılarda aynı kimilerinde farklı olarak belirlenmiş olması katılımcı öğretmenlerin aynı ürünü farklı şekillerde algıladıklarını göstermesi bakımından önemlidir. Büyük çoğunluğu öğretmenlik formasyonuna sahip katılımcıların bu yönde sergiledikleri tercih farklılıkları Wang vd. (2017) tarafından gerçekleştirilen araştırma verileriyle koşutluk göstermektedir. Uzmanlarca sağlanan puanlama eğitimini alan 20 puanlayıcı, yedinci sınıf öğrencilerince yazılmış 100 kompozisyonu kendilerine sağlanan çözümleyici ölçek aracılığıyla puanladıkları sırada yürütülen gözlemler ile Puanlayıcı Algısını belirlemeye yönelik formdan elde edilen veriler, puanlayıcı eğitimi veren uzmanlarla puanlayıcıların yanı sıra puanlayıcıların kendi aralarında birçok noktada farklı görüşlere sahip olduklarını ortaya koymuştur. Örneğin puanlayıcılar ile eğitim veren uzmanlar puanlaması en zor kompozisyonun seçiminde, puanlayıcı hatasına yol açan bölümler, yazılı ürünün odak noktası, metnin alımlanması ve fikirlerin düzenlenmesi konularında uyuşmazlıklar olduğu görülmüştür. Bu uyuşmazlıklar aynı ölçek kullanılsa bile bunun puanlayıcılar için yeterince açık ve anlaşılır olmaması durumunda farklı sonuçlara yol açabileceğini göstermesi bakımından önemlidir. Öte yandan ölçek ne kadar geçerli ve güvenilir olursa olsun puanlayıcının bilgisinin, birikiminin ve dikkatinin ölçeği algılamada ve kullanmada daha belirleyici olduğunu akla getirmektedir.

Araştırma bulgularımız arasında, değerlendirilen yazılı ürünün odaklanılan boyutlarının hem aşırı çeşitlilik göstermesi hem de aslında aynı kavramların farklı terimlerle ifade edilmesi puanlamayı yapan katılımcıların bu noktadaki birikimlerinin farklı olduğunu ortaya koymaktadır. Örneğin ölçekte dilbilgisi başlığı altında yer verilen bölüm sözdizimi, yazım ve noktalamayı içeren bir üst başlık olması gerekirken bu üç bileşenin yanında diğer bir bileşen olarak değerlendirilmiştir. Benzer biçimde dil kullanımı, bağdaşıklık, bağlaşıklık, bütünlük, tutarlılık ve akıcılık başlıkları altında değerlendirilen boyutlardan bazıları anlam bakımından belirsiz olup puanlayıcının örneğin tutarlılık, bütünlük ile bağdaşıklık ve bağıntı arasında fark görüp görmediği sorusunu akla getirmektedir. Kaldı ki dil kullanımı kavramı ile kastedilenin de yukarıda sıralanan boyutları kapsaması gerekirken ayrı bir başlık olarak değerlendirilmesi bu konuda puanlayıcıların farklı bakış açılarına ve algılara sahip olduğunu göstermektedir. Bu bulgu,

Wang vd. (2017)'nın araştırmasında varılan daha fazla örnek uygulama yapılması gerektiği yönündeki sonuca ek olarak, puanlayıcı eğitiminde yazılı anlatımın farklı boyutlarına ilişkin ayrımların daha kesin çizgilerle ortaya konması ve bunlar üzerinde özellikle durulması gerektiğini düşündürmektedir. Rahayu (2020)'nın yazma becerisinin değerlendirilmesi konusunda Endonezya'da ikinci dil olarak İngilizce öğreten 56 öğretmenle yaptığı çalışmada elde ettiği bulgular bu düşüncemizi güçlendirmektedir. Söz konusu araştırmada öğretmenlerden değerlendirme yöntem ve tekniklerine dair soruları içeren anketi cevaplamalarını ve ayrıca öyküleyici türde yazılmış olan iki kompozisyonu verilen çözümleyici ölçek ile puanlamalarını istemiştir. Dört bölümden oluşan anketteki sorular yazma becerisinin değerlendirilmesine yönelik bilgiyi, puanlama doğruluğunun etkililiğini, yazma becerisinin değerlendirilmesindeki seçimlerin etkililiğini ve yazma becerisinin değerlendirilmesindeki algılarını belirlemeye yönelik olarak hazırlanmıştır. Anketten elde edilen veriler öğretmenlerin, yazma becerisinin değerlendirilmesine ilişkin bilgilerinin, yazma becerisinin değerlendirilmesindeki seçimlerinin etkililiğinin, puanlama doğruluğunun etkililiğinin ve yazma becerisinin değerlendirilmesi uygulamasındaki algılarının; puanlamadaki başarılarını güvence altına almadığını göstermiştir. Diğer bir deyişle öğretmenlerin ankette verdikleri yanıtlar ile puanlama davranışları tutarlılık göstermemiştir. Öğretmenlerin yazma becerisinin değerlendirilmesine ilişkin bilgilerinin, yazma becerisinin değerlendirilmesindeki seçimlerinin etkililiğinin ve yazma becerisinin değerlendirilmesi uygulamasındaki algılarının artması puanlamalarını olumsuz etkilerken puanlama doğruluğunun etkililiğinin artması puanlamalarını olumlu etkilemiştir. Çalışma sonucunda öğretmenlerin puanlamadaki etkililiğinin değerlendirmedeki öğretmen kalitesini etkilediğini anlaşılmıştır.

Puanlayıcıların dilbilgisine ve yazılı ürünün biçimsel boyutuna daha fazla odaklandıklarını gösteren araştırma bulgularımız, alanyazında yaygın olarak yer alan verilerle uyumludur. Weigle ve Montee (2012)'nin yazma becerisinin bütünleşik bir yaklaşımla değerlendirilmesi sürecinde puanlayıcıların algılarını konu alan çalışmaları, puanlayıcıların yazılı ürünün biçimsel bileşenlerine farklı şekilde önem verdiklerini gösterirken yazma sürecinde öğrencilerin kullandığı alımlama tekniğine karşı da farklı tavırlar sergilediklerini ortaya çıkarmıştır.

Sonuç ve Öneriler

Yabancı dil öğretmenlerinin yazma becerisinin ölçülmesinde kullandıkları puanlama davranışlarının incelenmesinin amaçlandığı bu çalışma, araştırmaya veri sağlayan yabancı dil öğretmenleri olan puanlayıcıların yazılı anlatım ürününü puanlarken çok farklı davranışlar sergilediğini ortaya koymaktadır. Araştırma sınırlılıklarına rağmen alanyazında sorun olduğu belirtilen yazma becerisinin değerlendirilmesi sürecine yönelik önemli sonuçlar ortaya çıkarmıştır. Bu çerçevede yabancı dil öğretmenlerinin yazma becerisini ölçme amaçlı olarak ürettirilen bir ürünü puanlarken değerlendirme ölçütü kullanıp kullanmadığına yönelik sorunun bütünüyle olumlu bir biçimde yanıtlanamayacağı görülmektedir. Nitekim örneklem grubunda değerlendirme ölçütü kullananların sınırlı sayıda olduğu görülmektedir. Puanlayıcıların puanlama yaparken sergilemiş oldukları davranışların ne kadar çeşitlilik gösterdiğinin belirlenmesine ve ölçüt kullanıp

kullanmadıklarına yönelik farklı örneklem gruplarıyla çalışma yapılmasının, bu araştırmada elde edilen bu sonucun genelleştirilebilmesi için yararlı olacağı düşünülmektedir.

Öğretmenlerin yazma becerisini ölçme amaçlı olarak ürettirilen bir ürünü puanlarken ürünün hangi boyutlarına odaklandıklarını belirlemeye yönelik soru ise alanyazın verileriyle uyumlu olarak öğretmenlerin daha çok biçimsel boyuta odaklandıkları şeklinde karşılık bulmuştur. Oysaki yazma becerisinin değerlendirilmesinde tüm boyutların göz önünde bulundurulması gerektiği alanyazında pek çok çalışma tarafından önerilmiştir. Ek olarak, araştırmanın üç numaralı sorusuna yanıt olarak sınırlı sayıda puanlayıcının hata türleri arasında ayrım yaptığını ve hatanın ağırlığı ile puan arasında denklik kurduğunu gösteren veriler elde edilmiştir. Bu noktada öğretmenlerin yazma becerisinin değerlendirilmesi konusundaki bilgilerinin sınırlı olduğu düşünülmüştür. Dolayısıyla, öğretmen yetiştiren kurumların öğretim programlarında yazma becerisinin değerlendirilmesi ile ilgili derslere yer vermesinin gerekli olduğu görülmektedir. Ayrıca, hâli hazırda öğretmen olarak çalışanlar için çeşitli kurslar veya seminerler aracılığıyla bu konudaki bilgi eksikliğinin giderilmesi alan için önem arz etmektedir.

Son olarak yaş, cinsiyet ve deneyim ile puanlama davranışları arasında fark olup olmadığına yönelik araştırma sorusu kadınların erkek puanlayıcılara göre; kamu kurumlarında çalışanların özel kurumlarda çalışanlara göre; birden çok kademede ve üniversitede görev yapanların diğerlerine göre daha yüksek puanlar verdiği; buna karşılık deneyim ile verilen puanların yüksekliği arasında anlamlı bir ilişki olmadığı şeklinde karşılık bulmuştur. Bu farklılıkların daha farklı ve fazla örneklem grubunda da olup olmadığının anlaşılması ve bu farklılıkların altında yatan nedenlerin belirlenmeye çalışılması yazma becerisinde puanlayıcı etkisi konusundaki çalışmalara yeni boyut kazandırabilir. Puanlayıcı etkisine neden olan etmenlerin belirlenerek ortadan kaldırılmaya çalışılması yazma becerisinin değerlendirilmesi sürecinin daha güvenilir olmasını sağlar.

Kaynakça

Asassfeh, S. M. (2021). Holistic vs. analytic scoring between expository and narrative genres: Does the assessment type matter? *International Journal of Linguistics, Literature and Translation*, *4*(1), 215-220. https://doi.org/10.32996/ijllt.2021.4.1.21

Bachman, L. F. (2004). *Statistical analyses for language assessment book*. Cambridge University Press.

Baker, K. M. (2016). Peer review as a strategy for improving students' writing process. *Active Learning in Higher Education, 17*(3), 179–192. https://doi.org/10.1177/1469787416654794

Beck, S. W., Llosa, L., Black, K., & Anderson, A. T. G., (2018). From assessing to teaching writing: What teachers prioritize. *Assessing Writing, 37,* 68-77. https://doi.org/10.1016/j.asw.2018.03.003

Brown, J. D. (1989). Manoa writing placement examination. *Manoa Writing Board Technical Report*, 5.

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing*, *25*(3), 349–383. https://doi.org/10.1177/026553220809015.

Brown, G. T. L., Glasswell, K., & Harland, D. (2008). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, 9, 105-121. https://doi.org/10.1016/j.asw.2004.07.001

Calp, M. (2013). Serbest ve yaratıcı yazma tekniğine göre oluşturulan kompozisyonların yazılı anlatımın niteliği ve puanlama tekniği açısından karşılaştırılması. *Turkish Studies: International Periodical Fr the Languages, Literature and History of Turkish or Turkic, 8*(9), 879-898. https://doi.org/10.7827/turkishstudies/5340.

CECR (2018). *Cadre européen commun de référence pour les langues: Apprendre, enseigner, évaluer*.. www.coe.int/lang-cecr adresinden erişildi. Erişim tarihi: 15.05.2023

Cole, J. C., Haley, K. A., & Muenz, T. A. (1997). Written expression reviewed. *Research in the Schools, 4*(1), 17–34.

Coombe, C. A., Folse, K. S., & Hubley, N. J. (2007). *A practical guide to assessing English language learners*. University of Michigan.

Cooper, C. G. (1997). Holistic evaluation of writing. C. R. Cooper, & L. Odell (Yay. Haz.). *Evaluating writing* (ss. 3-33) içinde. National Council of Teachers of English.

Cooper, P. L. (1984*). The assessment of writing ability: A review of research*. Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00052.x

Crusan, D. (2010). *Assessment in the second language writing classroom*. University of Michigan.

Crusan, D. Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs and practices. *Assessing Writing*, 28, 43-56. https://doi.org/10.1016/j.asw.2006.03.001.

Çetin,B. (2002). *Kompozisyon tipi sınavlarda kompozisyonun biçimsel özelliklerinden kestirilen puanların anahtarla ve genel izlenimle puanlanmasından elde edilen puanlarla ilişkisi [The relation between scores predicted from structural features of an essay and scores based on scoring key and overall impression, in essay type examination]*Yayımlanmamış Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara.

Du, Y., & Wright, B. D. (1997). Measuring student writing abilities in a large-scale writing assessment. M. Wilson, Jr G. Engelhard, & K. Draney (Yay. Haz.). *Objective measurement: Theory into practice* (ss. 1-24) içinde. Abex.

Du, Y., Wright, B. D., & Brown, W. L. (1996). Differential facet functioning detection in direct writing assessment. In *Annual Conference of the American Educational Research Association 8-12 Nisan 1996* (ss. 1-21). ERIC.

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

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

Enginarlar, H. (1991). A quantitatıve and qualitative comparison of three techniques of grading ESL/EFL essays. *Journal of Human Sciences, 10*(1), 23-45. https://www.j-humansciences.com/ojs/index.php.IJHS/issue/view/27.pdf

Erman Aslanoğlu, A., & Şata, M. (2021). Examining the differential rater functioning in the process of assessing writing skills of middle school 7th grade students. *Participatory Educational Research, 8*(4), 239-252. https://doi.org/10.17275/per.21.88.8.4.

Ghalib, T. K., & Hattami, A. A. (2015). Holistic versus analytic evaluation of EFL writing: A case study. *English Language Teaching, 8*(7), 225-236. https://doi.org/10.5539/elt.v8n7p225

Göçer, A. (2011). Öğrencilerin yazılı anlatım çalışmalarının Türkçe öğretmenlerince değerlendirilmesi üzerine. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi, 30*(2), 71-97. https://doi.org/10.7822/egt34

Gyagenda, I. S., & Engelhard Jr, G. (2009). Using classical and modern measurement theories to explore rater, domain and gender influences on student writing ability. *Journal of Applied Measurement*, *10*(3), 225-246. https://d1wqtxts1xzle7.cloudfront.net/32596450/Gyagenda_Engelhard-libre.pdf

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, 8, 5-16. https://doi.org/10.1016/S1075-2935(02)00029-6

Han, T., & Huang, J. (2017). Examining the impact of scoring methods on the institutional EFL writing assessment: A Turkish perspective. *PASAA: Journal of Language Teaching and Learning in Thailand*, 53, 112-147. https://files. eric.ed.gov.tr/fulltext/EJ1153666.pdf

Hunter, K., & Docherty, P. (2011). Reducing variation in the assessment of student writing. *Assessment and Evaluation in Higher Education, 36*(1), 109-124. https://doi.org/10.1080/02602930903215842.

Jakobson, R. (1960). Linguistics and poetics. T. A. Sebeok (Yay. Haz.). *Style in language* (ss. 350-377) içinde. Mass. MIT.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485-505. https://doi.org/10.1177/0265532209340186

Kalay, S., & Büyükkarcı, K. (2020). English language teachers' views on teaching and assessment of writing skills. *SDU International Journal of Educational Studies, 7*(2), 262-286. https://doi.org/10.33710/sduijes.710062

Karatay, H. (2011). Süreç temelli yazma modelleri: Planlı yazma ve değerlendirme. M. Özbay (Yay. Haz.). *Yazma eğitimi* (ss. 21-43) içinde. Pegem Akademi.

Köksal, D. (2004). Assessing teacher' testing skills in ELT and enhancing their professional development through distance learning on the net. *Turkish Online Journal of Distance Education*, 5(*1*), 1- 11. https://dergipark.org.tr/en/pub/tojde/issue/16931/176755

Liu, L. (2022). Scoring judgment of pre-service EFL teachers: Does writing proficiency play a role?. *The Asia-Pacific Education Researcher, 31*(3), 333-343. https://doi.org/10.1007/s40299-021-00575-9

Li, J., & Huang, J. (2022). The impact of essay organization and overall quality on the holistic scoring of EFL writing: Perspectives from classroom English teachers and national writing raters. *Assessing Writing*, 51, 1-15. https://doi.org/10.1016/j.asw.2021.100604

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters?. *Language Testing, 19*(3), 246-276. https://doi.org/10.1191/0265532202lt230oa

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331-345. https://doi.org/10.1207/s15324818ame0304_3

Mede, E., & Atay, D. (2017). English language teachers' assessment literacy: The Turkish context. *Dil Dergisi, 168*(1), 43-60. https://dergipark.org.tr/tr/pub/dilder/issue/47674/602254

Mertler, C. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools, 12*(1), 101–113. https://doi.org/10.1177/1365480209105575

Mousavi, S. A. (2002). *An encyclopedic dictionary of language testing (3rd ed.)*. Tung Hua.

O'Malley, J. M., & Pierce, L. V. (1996). A*uthentic assessment for English language learners: Practical approaches for teachers*. Addison-Wesley.

O'Neill, P. (2011). Reframing reliability for writing assessment. *Journal of Writing Assessment, 4*(1), 1-15. https://escholarship.org/uc/item/6w87j2wp

Oruç, N. (1999). *Evaluating the reliability of two grading systems for writing assessment at Anadolu University preparatory school*. Yayınlanmamış Yüksek Lisans Tezi, Bilkent Üniversitesi, Ankara.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures and objective tests to evaluate ESL writing ability. *TESOL Quarterly, 17*(4), 651-671. https://doi.org/10.2307/3586618

Peterson, S., Childs, R., & Kennedy, K. (2004). Written feedback and scoring of sixth-grade girls' and boys' narrative and persuasive writing. *Assessing Writing, 9*(2), 160-180. https://doi.org/10.1016/j.asw.2004.07.002

Polat, M. (2003). *A study on developing a writing assessment profile for English preparatory program of Anadolu University School of Foreign Languages*, Yayınlanmamış Yüksek Lisans Tezi, Anadolu Üniversitesi, Eskişehir.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? Theory into Practice, 48, 4–11. https://doi.org/10.1080/00405840802577536

Rahayu, E. Y. (2020). The anonymous teachers' factors of assessing paragraph writing. *Journal of English for Academic and Specific Purposes*, *3*(1), 1-19. https://doi.org/10.18860/jeasp.v3i1.9208

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. A. J. Kunnan (Yay. Haz.). F*airness and validation in language assessment: Selected papers from the 19th language testing research colloquium* (ss. 129-151) içinde. Cambridge University.

Seviour, M. (2015). Assessing academic writing on a pre-sessional EAP course: Designing assessment which supports learning. *Journal of English for Academic Purposes*, 18, 84-89. https://doi.org/10.1016/j.jeap.2015.03.007

Smith, W. L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. M. M. Williamson & B. A. Huot (Yay. Haz.). V*alidating holistic scoring for writing assessment* (ss. 142-205) içinde. Hampton.

Stiggins, R. J., & Bridgeford, N. J. (1983). An analysis of published tests of writing proficiency. *Educational Measurement: Issues and Practices*, *2*(1), 6-19. https://doi.org/10.1111/j.1745-3992.1983.tb00679.x

Şeker, M. (2018). Intervention in teachers' differential scoring judgments in assessing L2 writing through communities of assessment practice. *Studies in Education Evaluation*, 59, 209-217. https://doi.org/10.1016/j.stueduc.2018.08.003

Thomas, N. (2020). Idea sharing: Are analytic assessment scales more appropriate than holistic assessment scales for L2 writing and speaking? *PASAA: Journal of Language Teaching and Learning in Thailand, 59*(1), 236-251. https://files. eric.ed.gov.tr/fulltext/EJ1239980.pdf

Tokur Üner, B., & Aşılıoğlu, B. (2022). İngilizce öğretiminde ölçme ve değerlendirme sürecine ilişkin öğretmen görüşleri. *EKEV Akademi Dergisi*, 89, 25-50. https://dergipark.org.tr/en/pub/sosekev/issue/71371/1147452

Turgut, M. F. (1990). E*ğitimde ölçme ve değerlendirme metotları*. Saydam.

Wang, J., Engelhard Jr, G., Raczynskia, K., Song, T., & Wolfec, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing*, 33, 36-47. https://doi.org/10.1016/j.asw.2017.03.003

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, *16*(3), 194-209. https://doi.org/10.1016/j.jslw.2007.07.004.

Weigle, S. C., &Montee, M. (2012). Raters' perceptions of textual borrowing in integrated writing tasks. *Studies in Writing*, 27, 117–152. https://doi.org/10.1163/9789004248489_007

White, E. (1994). Issues and problems in writing assessment. *Assessing Writing*, *1*, 11–27. https://doi.org/10.1016/1075-2935(94)90003-5

Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing*, *27*, 11-23. https://doi.org/10.1016/j.asw.2015.06.003

Wind, S. A., & Engelhard Jr, G. (2012). Examining rating quality in writing assessment: Rater agreement, error, and accuracy. *Journal of Applied Measurement*, *13*(4), 1-15. https://d1wqtxts1xzle7.cloudfront.net/31697482/SW_GE_2012-libre.pdf (d1wqtxts1xzle7.cloudfront.net)

Wiseman, C. S. (2012). A comparison of performance of analytic vs holistic scoring rubrics to assess L2 writing. *Iranian Journal of Language Testing, 2*(1), 59-92. https://www.ijlt.ir/article_ 114361_9544f0e7ef140d3731098f945f34a848.pdf

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

Zorbaz, K. Z. (2013). Yazılı anlatımın puanlanması. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *13*(1) 179-192. https://openaccess.mku.edu.tr/xmlui/bitstream/handle/20.500.12483/1993/Zorbaz%2c%2Kemal%20Zeki%202013.pdf?sequence=1&isAllowed=y