An Overview of Writing Rubrics in Doctoral Dissertations in Turkey

Gökhan Arı[i]
Bursa Uludağ University

Abstract

Writing rubrics have been used in doctoral dissertations in Turkey to assess student writing for nearly twenty years. This study aims to determine which features are assessed in the rubrics used in doctoral dissertations. Twenty-five rubrics were selected to determine the analysis of validity and reliability, rubric dimensions, features of criterion descriptions, and performance levels. The results showed that researchers did not attach much importance to the validity of rubrics, and some studies provided insufficient information about reliability. Rubrics typically have three and five performance levels. As a result of the content analysis based on the descriptions, nine categories (genre-oriented content, organization, style, presentation, mechanics, vocabulary, sentences, textuality, process) and 104 features were identified. The frequency of these features was 567. According to the emerging categorization in the present study, we observed that the rubrics used in doctoral dissertations were similar to the criteria in the 6 +1 Trait Writing Model. However, the descriptions were found to be more straightforward and concise. The features in the rubrics were similar to one another yet differed depending on the purpose, type, and length of descriptions. Some of the rubrics had descriptions that were ill-suited to the nature of the rubric. The results were discussed, presenting recommendations.

Keywords: Writing Assessment, Writing Rubrics, Features Of Criteria

DOI: 10.29329/ijpe.2020.332.24

---

[i] **Gökhan Arı,** Assoc. Prof. Dr., Turkish Education, Bursa Uludağ University, ORCID: 0000-0001-7054-2209

Email: gokhanari@uludag.edu.tr

INTRODUCTION

Studies on writing education appear to have begun late due to the effect of matters such as the relatively late start of schooling in Turkey (Göğüş, 1971; Göğüş & Yücesan, 1988), a lack of emphasis being placed on mastering first language skills (Göğüş, 1978; Özdemir, 1983; Şimşek, 1983), an underdeveloped writing tradition, and a focus on reading more than writing (Çakın, 1966; Göğüş, 1971; Şengiz, 1976). Graduate education gained importance upon the establishment of the Council of Higher Education (CoHE) in 1981, the increase of universities in number, the transformation of teacher training schools to faculties, and the establishment of institutes (graduate schools). Thus, graduate studies on writing education started in the late 1980s due to these reasons causing direct or indirect impacts.

The first assessment tool used in graduate studies was in a master's thesis in 1989. Translated to Turkish by Erdiken (1989), the tool was indeed an adaptation from a rubric. It was designed for assessing the writing of deaf students. It was not like a rubric in form but a tool that had a scale of scores. In his doctoral study in 1996, Erdiken assessed five traits (content, organization, vocabulary, grammar, mechanics) showing evaluative scales (e.g., Excellent-Good-Fair-Poor) in his analytical rubric. The second rubric was used in a doctoral dissertation in 1990. This holistic rubric (Enginarlar, 1990) included eight features and a 5-point rating scale, but there were no descriptions, and the rubric consisted of keywords. In the study, Turkish high school students' writings in Turkish and English were assessed. In this sense, the first assessment tools identified were, in fact, rubrics. Rubrics were used in four master's theses written in the departments of English and German Education between 1989 and 1992 (I did not include these rubrics here because they did not evaluate Turkish writing or were used to assess the writing of disabled students).

The first assessment tool assessing writing in Turkish was developed by Sever (1993). This tool was a scoring scale, not a rubric. The scoring scales were also used in eight theses written from 1993 to 2002 (Sever, 1993; Pehlivan, 1994; Özbay, 1995; Duman, 1997; Şimşek, 2000; Deniz, 2000; Temur, 2001; Tekşan, 2001). The first rubric prepared for the assessment of writing in Turkish was used in 2002.

The criteria of a scoring tool used in writing assessment vary according to raters' purpose and experience. Although a norm-based assessment (Sezer, 2005) is dominating in the Turkish education system, there seems to be a lack of standardization in writing assessment (Arı, 2008). While the features measured in many scoring tools are similar to one another (e.g., 6 + 1 Trait Writing model; Arı, 2015; Coşkun, 2005; Glass, 2005; Özbay, 1995; Rankin, 2015; Sarıca, & Usluel, 2016; Sever, 1993; Sezer, 2005), different criteria sets are also deployed. However, there is no evidence of studies examining the writing rubrics used in scientific research in Turkey to determine which features are assessed in students' writing. The features identified in this study are essential in forming a theoretical background for a standardized writing assessment.

Writing rubrics

Many experts agree that accurate and formative assessment of student writing is challenging (Graham, Hebert, & Harris 2011; Humphry, & Heldsinger, 2014; Huot, 2002, Schoonen, 2005; Reddy & Andrade 2010, etc.) Student writing is evaluated using one of the approaches, including checklist approach, rating scale approach, memory approach, and descriptive approach (Brualdi, 2002). Rubrics are descriptive; therefore, they can be one of the most appropriate tools for writing assessment. A holistic or analytical rubric has three essential features: assessment criteria, quality description, and a scoring strategy (Popham, 1997, p. 72). The central inquiry of this study includes examining which features are contained in quality definitions of writing rubrics.

Flynn, Tenam-Zemach, and Burns (2015) claim that rubrics serve as an essential and accountable tool for the standardization of education, and the use of rubrics in education has been

ground-breaking. Rubrics stand out as "instructional illuminators" (Popham, 2000, p. 75; Wilson, 2006) in terms of their use for teaching and assessment purposes (Gunning, 2006; Martin-Kniep, 2000; Moskal & Leydens, 2000; Popham, 1997; Wilson, 2006). In this respect, students tend to consider rubrics for learning purposes while teachers see them as an assessment tool (Li, & Lindsey, 2015). Rubrics generally assess language use, style, meaning-making, and organizational skills in students' writings. For the scoring strategy, rubrics are divided into holistic and analytic rubrics. Teachers frequently apply holistic rubrics as they are easy to prepare and assess (Babin, & Harrison, 1999; Martin-Kniep, 2005). Analytical rubrics are more practical to assess each feature individually (Faigley et al., 1985; Crehan, 1997). Nevertheless, it is still controversial which approach is more suitable for assessing student writing (Weigle, 2002). Some experts criticize the holistic rubric (White, 1985; Elbow, 2000) and holistic scoring for various reasons.

In assessments where rubrics are used, raters cannot handle each criterion independently, and the emerging halo effect may lead to severe problems (Humphry & Heldsinger, 2014). In other words, raters both try to abide by rubrics and make biased decisions by being affected by other features (Elbow, 2000). In this regard, the criteria, the quality definitions, and the areas or boundaries of these definitions are significant for raters during decision-making process (raters' knowledge and experience are also critical).

Text production requires complex processes; thus, assessing texts with rubrics and achieving inter-rater agreement are crucial and quite challenging (Beyreli & Arı, 2009). In many studies, there is little agreement reported among raters (Graham, Hebert, & Harris, 2011). Examining the quality, validity, and reliability of rubrics used in higher education and the state of usage of formative assessment, Reddy and Andrade (2010) revealed some deficiencies. Both designing and evaluating a writing rubric for scientific research and ensuring inter-rater agreement and thus validity and reliability requires well-management of these problematic processes and an effort to reach perfection.

The criteria to be included in a rubric vary with respect to the researcher or teacher's purpose. Knoch (2011) states that there is no apparent unity in assessment approaches, yet grammatical, lexical, and syntactic features are categorically similar, even though their feature descriptions are different, which should be regarded as normal. Some studies produced complex findings in the assessment of writing for different tasks (Knoch, 2009). Put differently, the criteria in rubrics may differ based on the characteristics of writing purposes, tasks, and types. Graham, Hebert, and Harris (2011) conducted a meta-analysis to determine the traits to be rated in a typical writing assessment, including vocabulary, convention, setting, characters, setting, style, organization, detail, spelling, communication, sentence fluency, theme, usage, and place. In another meta-analysis by Graham, Hebert, and Harris (2015), they found that experimental studies statistically enhanced writing quality whereas the implementation of the 6 +1 Trait Writing Model did not have a meaningful effect on the development of students' writing.

In the US and many other countries, 6 or 6 + 1 Trait Rubrics are used in large-scale assessments (Grabe & Kaplan, 1996). These rubrics initially developed by Diederich, French, and Carlton (1961) assess ideas and content, organization, voice, word choice, sentence fluency, and convention traits (Roid, 1994; Spandel, 1996; Jakobson, 2005).

Many of the rubrics used for writing assessment in the graduate research of the last 20 years in Turkey, unsurprisingly, considered the abovementioned traits as dimensions or criteria. In assessing writing, similar traits can be taken into account, depending on the purpose and type of assessment. However, the features included in rubrics and the extent to which they are included are essential for standard assessment or assessment close to standard. However, there is no widely used, conventional, or standardized teaching and measurement tool or rubric in Turkish schools (Arı, 2008). Such tools are used in scientific research prepared to assess students' writings. Therefore, the traits in rubrics vary with respect to the purpose of research.

This study aims to determine the validity and reliability of the rubrics used in doctoral dissertations, the rubric dimensions, the features of students' writings assessed with these rubrics, and the similarity and quality of criterion descriptions. Thus, the present study can provide future researchers who will use rubrics in writing assessment to categorize frequently used criteria and items and also guidance about what to consider in terms of validity and reliability. The research questions about the rubrics used in writing assessment in doctoral dissertations were as follows;

1. What analyses are undertaken for the validity and reliability of the writing rubrics?
2. What are the dimensions and criteria in the writing rubrics?
3. What features are included in the descriptions of the writing rubrics?
4. Are there similarities in the descriptions of the writing rubrics?
5. Are there any incorrect descriptions in the writing rubrics?

METHODOLOGY

When considering the nature of the research objective and the material examined in this study, content analysis was used with an inductive approach to determine the focal point of the descriptions in the rubrics. For data collection, document analysis was employed while content and category analyses were conducted to analyze the data.

Materials and Procedure

This section included the information about (i) how the materials (rubrics) were accessed, (ii) what features of the materials led to their exclusion from our analysis, (iii) the description of general characteristics of the studies included in the study, (iv) how criteria descriptions of the rubrics were coded, (v) the coding process, and (vi) the coder reliability.

In the database of the YÖK (CoHE), the use of keywords 'writing (yazma), writing education (yazma eğitimi), written expression (yazılı anlatım), composition (kompozisyon), assessment of composition (kompozisyon değerlendirme), assessment of written expression (yazılı anlatımı değerlendirme), scoring writing (yazmayı puanlama), writing assessment (yazmayı değerlendirme), writing rubric (yazma rubriği), scoring rubric in writing (yazmada puanlama rubriği), measurement of writing (yazmayı ölçme), scoring key for writing (yazmayı puanlama anahtarı), scoring writing (yazma puanlama), and writing scale (yazma ölçeği)' yielded 141 graduate studies where writing assessment was conducted in Turkish. Sixteen of the studies examined the writings of deaf or mentally disabled students and students with learning difficulties. Fourteen of them studied the writings of students learning Turkish as a foreign language, and four were about bilingual writing. 103 of them investigated the writings of students speaking Turkish as a native language.

Of 103 graduate studies, 45 of them were doctoral dissertations while 58 of them were master's theses. Thirty-one doctoral dissertations and 39 master's theses used rubrics as an assessment tool. The scoring scale was employed in 14 doctoral dissertations and 19 master's theses. Sever's (1993) scale was utilized in 7 doctoral dissertations and six master's theses. Besides, five master's theses used the scale recommended in the national curriculum (MEB, 2006). Many scoring scales had insufficient information about validity and reliability or contained methodological issues. Most studies did not perform an inter-rater agreement analysis. Either there was no information about the agreement, or the researcher of the study conducted the analysis by her/himself. Besides, most of these tools listed only keywords (e.g., introduction, sentences, main idea) as scoring criteria and presented the assigned points. It was not at all clear what was assessed with these items. For these reasons, graduate studies containing scoring scales were excluded from the present analysis.

Most of the master's theses in which rubrics were used inherited ambiguities. Some scoring scales were converted into rubric-like tools MEB (2006) by *grading* the scales. Some of them used the rubrics in doctoral dissertations. Some rubrics were called “analytical," even though they were actually "holistic." Some of those called “rubric” were, indeed, a checklist. One of the studies provided information about the preparation procedure of the rubric, though the rubric itself was not displayed in the study. Also, most studies did not include validity and reliability analyses. For these reasons, master's theses using rubrics were excluded from the analysis. Reddy and Andrade (2010) excluded master's theses for almost the same reasons. As a result of the inclusion and exclusion process, a total of 25 rubrics in 24 doctoral dissertations were identified for examination.

Table 1 displays the thesis numbers and years of doctoral dissertations in which rubrics were used.

Table 1. Rubric Number and Year

Year	Number	Year	Number	Year	Number	Year	Number
2001	-	2006	1	2011	1	2016	3
2002	1	2007	1	2012	1	2017	6
2003	-	2008	2	2013	1	2018	2
2004	-	2009	1	2014	2	2019	-
2005	-	2010	1	2015	1	2020	1

To present general information about the material, we categorized the research type of the graduate studies where rubrics were utilized, the text type assessed with rubric, the grade and degree of the participating groups, and the type of rubric.

Of the reviewed studies, 16 (64%) were experimental studies, 5 were case studies (20%), and 4 (16%) were action research. 2 of the rubrics were adapted rubrics, and 2 of them were rubrics that were recommended in books. 21 rubrics were designed and developed by the researchers themselves. 12 (48%) of the rubrics were used to assess compositions (general writing), 8 (32%) for narration, 2 (8%) for informative writing, 2 (8%) for argumentative writing, and 1(4%) for persuasive writing. 7 (28%) of the rubrics were used in primary school, 15 (60%) in secondary school, and 3 (12%) in undergraduate student writing. 3 (12%) of the rubrics were holistic, while 22 (88%) were analytical.

After this categorization, we initiated the data collection procedure and randomly assigned codes to the selected dissertations such as D1, D2, D3, etc. In the first stage, we coded and listed the rubrics with MS excel in terms of validity and reliability analyses, the sample numbers used in the assessment, the grades/degrees of the sample in the assessment, the number of raters, information about the pilot application, and the number of writing assessed for inter-rater agreement. Some studies that provided information about the inter-rater agreement were excluded from the analysis since they inconsistently presented the calculations (some of them calculated dimensions individually or provided a single score while some others had an insufficient number of writing for the agreement analysis, etc.). These cases were recorded, yet it was simply not possible to present them coherently. Nevertheless, we took notes of the agreement analyses for the interpretation of the findings.

In the second stage, after dimensions in rubrics, sub-dimensions/criteria, features, levels, and the number of features evaluated in level descriptions were listed via MS Excel, the criteria descriptions in the rubrics were encoded. This stage can be considered as a draft. The preliminary study aimed to determine the possible categories relating to particular codes or the relations of these categories to one another. In coding, the concepts in the description of the excellent level in the rating scale were taken into consideration, but the descriptions of the rest of the performance levels were also examined. As the results showed, some terms in the excellent level overlapped with the terms from a lower performance level. However, the features of an examined criterion were coded in a disorderly manner due to the differences in terms, ambiguities in descriptions, and synonyms in rubrics. Besides, coding errors (different codes with the same concept, inaccurate counts, code inconsistencies, etc.) and

inconsistencies with code names were identified. In coding, the differences in terms and ambiguities in the criteria descriptions based on the rating scale caused challenges. We sought and found consistent solutions to minimize the inconsistencies with the codes and terms and ambiguities in descriptions taking three weeks away from the study. A qualitative research expert was consulted in this regard. The results demonstrated that it was essential to establish a focal point in the criteria descriptions. There were still problems in coding the criteria and features. We wrote a more general code for some criteria and features. In her taxonomy, Knoch (2009) excluded the raters' subjective characteristics and the features related to the writer that affect the writing product. Although we firstly preferred the exclusion of the features with subjective judgments involved, they were still encoded. The reason was that in some rubrics, the number of rater decisions based on subjective judgments was substantially high in number. Disregarding this type of rubrics may be against the nature of assessment with writing rubrics and disrupt the integrity of the study.

After a three-week break, the codes in the draft list were set aside, and the rubrics were converted into texts. These documents were transferred to MAXQDA Analytics Pro 2020 (Release 20.2.0) software. The solutions discovered during the draft preparation were utilized, which resulted in returning to the beginning of coding. During the coding process, no category or subcategory emerged. After the coding was completed, the codes in all rubrics were reviewed. We had found eight overlooked traits before the completion of coding. Besides, inconsistencies or errors in performance level descriptions were noted. The rubrics, converted documents, criteria, and codes were double-checked. A limited number of codes were reorganized. At the last stage, we identified the categories from the codes through the software based on the literature and the patterns in the rubrics. At this stage, the qualitative research expert helped brainstorming about the coding process. Some codes did not develop into meaningful categories, but they were kept close to similar categories. A literature review was conducted to determine the state of the two codes that could be evaluated under the same category. After the codes and documents were reviewed, the development of the categories was complete. Reviewing the codes under the categories, we found that some of these codes allowed the construction of subcategories.

Reliability

A doctoral student (research assistant), whose research was on writing assessment, was asked to assist the analysis process as a second coder to achieve reliability. A list was created consisting of the codes, the full and short descriptions of the codes, the state of their use and non-use, and examples to determine inter-rater agreement, which was exemplified in a rating session. In a different rating session, the rationale for the coding process was explained, and clear and unclear descriptions were illustrated using different rubrics. We showed an example of how to encode all the descriptions in a rubric. In another rating session, the research assistant was asked to code a different rubric entirely. Then, the codes of the two raters were negotiated. Upon completing this stage, the second rater was asked to encode the features in the rubrics. Kappa analysis was performed for inter-rater reliability. Table 2 shows the findings.

Table 2. Inter-rater Agreement

		Coder1		
		1	**0**	
Coder 2	**1**	a = 539	b = 28	567
	0	c = 0	0	0
		539	28	567

Table 2 displays that 539 of the 567 codes were consistent, while 28 of them had inconsistencies.

The Kappa analysis was conducted based on the data in the table, which revealed a near-perfect agreement with a Kappa value of 0.95. After the negotiations in two rating sessions, which lasted three hours 19 minutes, we reached a consensus regarding the 28 inconsistent codes. In light of this consensus, the codes and categories in the rest of the rubrics were re-evaluated, and the data analysis was finalized.

Due to our illustrations of several scientific shortcomings and errors in the rubrics and hence the doctoral dissertations, the code names (D1, D2, D3, etc. as document numbers) were adopted instead of the names of the studies for ethical considerations.

FINDINGS

In this section, the findings were displayed, interpreted, and illustrated using tables based on the targeted research questions in the current study.

Validity and reliability of the rubrics in the studies

This section shows a selection of validity and reliability analyses of the 25 rubrics examined to address the first research question.

In all doctoral dissertations examined, the researchers stated that they sought expert opinion before or during the rubric preparation. Among the experts were academicians, while teachers were consulted eminently. In a small number of studies, researchers stated to confer with doctoral students about their opinions. In D4, D11, and D17, detailed information was provided concerning the alterations in the rubrics following the expert opinions. The rest of the studies noted, "alterations were made in the rubric for its final form."

Table 3 illustrates whether or not a pilot study was conducted, the type of validity and reliability analyses (if any), the number of raters, and the number of writing assessed for inter-rater agreement in developing rubrics.

Table 3. Information on validity and reliability analyses of rubrics

doc	Pilot	Validity	Reliability	Rater n	Paper n
D1	-	Lawshe	Correlation (Pearson)	2	10
D2	+	-	Correlation (Pearson)	6	139
D3	+	-	Correlation (Pearson, Freidman, Kendall's w),	6	200
D4	-	-	Correlation (?)	3	50
D5	-	-	Correlation (?)	3	20
D6	-	-	-	3	515
D7	-		Correlation (Pearson)	3	154
D8	-		Correlation (?)	2	68
D9	+	-	Similarity (Miles-Huberman)	2	25
D10	+	-	Correlation (Spearman-Brown, Kendall's w)	3	50
D11	-		Correlation (?)	4	10
D12	-		correlation (Kendall' w)	3	48
D13	+	Lawshe	correlation (Kendall' w)	6	21
D14	-	-	-	3	(?)
D15	+	-	Correlation (?)	2	(?)
D16a	-	-	Correlation (?)	3	30
D16b	-	-	Correlation (?)	3	30
D17	-	-	-	4	(?)
D18	-	-	Correlation; Repeated measurement	(1)	174
D19	-	-	Correlation (?)	25	1

D20	-	-	-	3	(?)
D21	+	-	Correlation (?)	2	322
D22	-	-	Similarity (Miles-Huberman)	5	(?)
D23	-	-	-	2	(?)
D24	-	Lawshe	correlation (Fleiss Kappa)	5	80

(?): No information about the concept was delivered in the study.

As presented in Table 3, the seven studies (D2, D3, D9, D10, D13, D15, and D21) utilized writing samples of students before or during the rubric preparation. Only two of the studies (D3, D21) in which the pilot study was conducted described the process in detail. In D3, the rubric first described three performance levels and later five levels. Besides, in the first pilot application, it was insufficient to score the quality, while in the second, the measurement results were inconsistent. The study described four levels at the last stage, and it was decided to use the rubric. D21 reported that the pilot application evidenced that some descriptions were insufficient and corrected afterward.

After consultation with an expert for the rubrics, only three studies (D1, D13, D24) used the Lawshe technique and computed content validity. 22 (88%) rubrics did not perform the analysis of content validity. In other words, the attempt was to ensure internal consistency with the rubrics.

In five studies (D6, D14, D17, D20, D23), no validity and reliability analyses were performed (see Table 3). In D6, the writings of 515 students were graded by three raters, resulting in an average score. In D23, an experimental study with 70 students, no information was provided about the number of writings, yet two raters graded the writings and gave an average score. In D14, D17, and D20, the information about validity and reliability was strikingly insufficient.

Table 3 shows that a reliability analysis was performed for 20 rubrics. In D9 and D22, the similarity ratio was calculated with the Miles and Huberman formula. This technique is not quite commonly used in such studies. In the remaining 18 rubrics, the inter-rater agreement was determined with correlation analysis. The information about correlation was missing in 10 of these studies. D18 did not report the type of correlation analysis performed and the number of raters. It appeared that the researcher conducted two correlation tests using the test-retest technique. Pearson correlation analysis was used in 4 studies (D1, D2, D3, D7), Kendall's w (D3, D10, D12, D13) in four studies, Friedman (D3) in one study, Spearman-Brown (D10) in one study, and Fleiss Kappa in one study (D24). In D3, where more than one correlation was used, the inter-rater correlation coefficients were calculated with Pearson. At the same time, the significance level was measured with Friedman and Kendall's W. In D10, the split-half reliability was estimated with the Spearman-Brown formula and the agreement with Kendall's W.

As presented in Table 3, three raters participated in nine studies, two in five studies, six in three studies, two in two studies, two in two studies, one (D18) in one study, and 25 in one study. In D19, an experimental study, it was noteworthy that 25 rated assessed one paper for inter-rater agreement. Although the studies provided information about the raters' professions and their years of work experience, there was no information about their experience in writing assessment or rater training and rating sessions. Only D3 reported that the raters were given a two-session training.

Although we checked multiple times, we did not encounter the number of writings assessed in six studies with reliability analysis (D14, D15, D17, D20, D22, D23). Among these, D15 and D22 performed reliability analysis. The credibility of the technique used in D22 was disputable. No validity and reliability analyses were conducted in D20. Besides, it was noted that three raters assessed the writings used in the pre-test and post-test. No information was found about validity and reliability in D14 and D17. The results of expert opinion were unclear. The function of the scores of raters was ambiguous in these studies.

D1, D13, and D24 were the only studies that explicitly expressed the process of validity and reliability analyses, and they seemed to obtain their statistical results following their selected methods.

Among the three studies, we can argue that the number of writings (10) assessed in D1 was relatively low.

Dimensions and the number of criteria in the rubrics

In this section, we aimed to answer the second research question where we included descriptive information about the performance levels of the rubrics, the presence/absence of level descriptions, the dimensions included in the rubrics, the number of criteria in the dimensions, as well as the findings of the number of description-driven codes/features. Table 3 presents the concerning information.

Table 4. Dimensions in rubrics, the number of criteria and description-driven codes

doc	level	Description	Dimension n	Dimensions/number of criteria in dimensions	Number of criteria	*Number of code*
D1	3	+	3	form (2), language and expression (13), textual elements of narrative (11)	26	*35*
D2	6	+	4	structure (4), plot (3), organization (3), spelling and punctuation (3)	13	*19*
D3	4	+	3	external structure (2), language and expression (4), organization (4)	10	*40*
D4	5	-	3	story elements (6), formal features of a story (3), language and expression (1)	10	*40*
D5	5	-	3	form (5), language and expression (15), spelling and punctuation (7)	27	*24*
D6	4	+	4	outline and page layout (2), narration (10), agreement (2), spelling and punctuation (2)	16	*22*
D7	4	+	2	sentence (6), paragraph (5)	11	*11*
D8	3	+	-	-	9	*13*
D9	3	+	-	-	7	*10*
D10	4	+	2	quality of writing (5), cognitive design procedures (11)	16	*23*
D11	4	+	4	external structure (2), language and expression (4), plan (4), spelling and punctuation (1)	11	*25*
D12	3	+	-	-	12	*17*
D13	4	+	4	plot (4), fiction (5), credibility (3), narration (2), external structure (2)	16	*19*
D14	3	+	3	organization (4), language and expression (7), form (3)	14	*22*
D15	5	+	-	-	7	*23*
D16 a	3	+	4	page layout and title (5), spelling and punctuation (2), plan (12), language and expression (6)	25	*25*
D16 b	3	+	4	page layout and title (5), spelling and punctuation (2), plan (12), language and expression (6)	25	*25*
D17	5	+	-	-	8	*22*
D18	5	-	6	content (6), narration (7), organization (7), sentence structure, (3), vocabulary (3), punctuation and grammar (6)	32	*19*
D19	3	+	5	content (7), coherence (8), vocabulary (4), sentence structure (6), mechanics (4)	29	*22*
D20	5	-	3	creativity (5), plan (8), narration (12)	25	*19*
D21	5	+	-	-	5	*7*
D22	5	+	-	-	7	*9*
D23	5	+	7	ideas (6), organization (6), vocabulary (6), voice (5), sentence fluency (5), conventions (6), presentation (5)	39	*40*
D24	5	+	6	content (5), narration (7), organization (5), sentence structure (5), vocabulary (5), punctuation and grammar (7)	34	*36*

According to Table 4, 10 (40%) of the rubrics had five levels, eight (32%) had three levels, six (24%) had four levels, and one of them (4%) had six levels. However, although D4, D5, D18, and D20 placed these levels in the rubrics, the characteristics or descriptions of the levels were missing. D17 and D23 had five levels, but only the first, third, and fifth levels were defined, and no description was provided for the second and fourth levels.

Seven of the rubrics (D8, D9, D12, D15, D17, D21, D22) did not have dimensions, and the criteria were listed without following any hierarchical order. Six of the rubrics defined three dimensions, and another six had four dimensions. Two rubrics had two dimensions. Another two constructed six dimensions. One rubric developed five dimensions, and one study had a rubric with seven dimensions.

Examining the rubrics, we found out that 10 dimensions for organization, 10 dimensions for mechanics, nine dimensions for presentation, and eight dimensions for language and expression were frequently employed in naming practices of dimensions. The less common dimensions were expression (5), sentence (5), content (4), and vocabulary (4). The dimensions seemed to be similar in general. Some rubrics (e.g., D1, D4, D13) marked the genre-based naming practice. The naming of dimensions in some studies (e.g., D7 and D10) signaled that some unique cases were measured. Nevertheless, it was not the content that was unique but the naming of dimensions. There were two dimensions (sentence and paragraph) in D7, but the criteria included spelling and punctuation, vocabulary, parts of writing (introduction, body, conclusion) for the assessment of writing. Looking at the names of the dimensions in D10, one can assume that the rubric measured the process; however, as in other rubrics, it included the features of genre-based content, organization, etc. as the criteria.

As Table 4 shows, D23 (39) was the rubric with the highest number of criteria, while D21 (5) used a rubric where we found the lowest number of criteria. When the criterion descriptions were coded, the rubrics D3, D4, and D23 had the highest number of codes/features (40 each). Although there were 10 criteria each in D3 and D4, we developed many codes. The most important reason was that there were many terms in the criteria and lengthy sentences and overly broad meanings in the descriptions. On the other hand, although some rubrics consisted of a large number of criteria, the number of codes was few since some of the codes described the same feature (e.g., D19, D20). D21 (5) used a rubric where we defined the lowest number of codes. D21 (the rubric was called "Analytically Graded Scorecard for Argumentative Text Writing" although it was really a holistic rubric) and D22 were holistic rubrics, and both rubrics had a small number of criteria. Although D2 used a holistic rubric, the number of criteria was 13.

Features of Criteria Descriptions

This section presents the encoded and categorized features in the descriptions of the rubrics used in the doctoral dissertations in Turkey. Table 5 shows the features of the rubrics for writing assessment, including the frequency scores and categories generated based on these scores.

Table 5. Category, subcategory, and code feature for writing assessment

Category/traits	*Subcategory:* codes/features
(genre-based) Content (158)	*Narrative [60]: Story elements (56)*: characters (19), place (9), time (9), problem (6), initiating event (2), reaction (4), events (4), nature of the event (4), plot (3). U*se of story techniques (4):* dialogue (2), inner speech (2)
	Information (48): *providing detailed information (13), main idea (11), topic (9), supporting ideas (6), commenting (4), justification (3), conveying a message (2)*
	Discussion (14): supporting a claim (4), defended claim (3), counterclaim (2), grounds of a counterclaim (2), repetition of a claim (1), rejection of a counterclaim (1), problem (1)
	Style of expression (12): use of narrative styles (5), explanations (4), describing (3)
	Idea development (12): the use of techniques for idea development (5), exemplification (4), quotation (2), describing (1)
	Content-appropriateness for text type (12*)*.
ORGANIZATION (131)	*Textual parts (64):* conclusion (17), title (16), introduction (15), body (14), textual parts (1), subtitle (1).
	Integrity (27): topical integrity (18), paragraph integrity (9)
	Connection (26): Connection among paragraphs (10), connection among sentences (9), connection among ideas (3), connection among events (3), connection among the parts of writing (1)
	Train of thought (13): the order of events (6), the order of ideas (5), the order of importance (2)

	Clarity of organization (1)
STYLE (64)	fluency (10), originality (9), perspective (6), rhetoric (3), audience orientation (5), mood (4), attractiveness (3), persuasiveness (3), creativity (3), uniqueness (2), risk-taking (2), naturalness (3), neatness (2), clarity (2), simplicity (2), sensation (1), amusement (1), flexibility (1), reality (1), selectivity (1)
PRESENTATION (59)	*Spacing (24):* Margins (8), line spacing (8), paragraph spacing (6), word spacing (2)
	Legibility (13), indenting (11), line quality (3), writing style (7), use of visual elements (1)
MECHANICS (59)	Spelling (20), punctuation (18), grammar (15), standard language use (4), language use (2)
VOCABULARY (44)	word senses/word selection (19), word diversity (8), avoiding word repetition (8), use of phrases (8), use of keyword (1)
SENTENCES (36)	*Sentence diversity (14)*: Cause-effect (4), goal attainment (4), descriptive sentences (2), ellipsis (2)
	Sentence structure (10): sentence formulation (4), different sentence structure (6)
	sentence meaning (6), different sentence length (5), different syntax (3)
TEXTUALITY (10)	coherence (6), cohesion (4).
PROCESS (5)	Drafting (2), method (1), editing (1), sharing (1), task (1).
TOTAL 9 (567 frequencies)	*11 subcategories and 104 different codes*

In the analysis, a total of 104 different codes/features emerged, comprising of 32 codes and five subcategories in genre-based content, 17 codes in five subcategories in organization, 20 codes in style, nine codes and one subcategory in presentation, five codes in mechanics, five codes in vocabulary, nine codes and two subcategories in sentence, two codes in textuality, five codes in process. The total code frequency was 567. No subcategories were generated from the categories of style, mechanics, vocabulary, textuality, and process.

Genre-based content appeared to be the category with the highest number of codes. This particular category included the subcategories of narrative, information, discussion, style of expression, idea development, and content-appropriateness for text type. The subcategory with the highest frequency was narrative (60), and the most frequent code was characters (19). Writing tasks and topics assessed by the rubrics were not context-based but generally depended on scoring genre-based student writing. Since nearly half of the rubrics were prepared to assess general writing (composition), they also included the main idea, supporting ideas, topic, providing detailed information, and idea development as well as narrative features such as characters and events or a general feature including content-appropriateness for text type in the performance descriptions. Although those rubrics that assessed narrative texts focused on the features regarding story elements, some rubrics also included the subcategories of the main idea, topic, and message. The descriptions reserved for argumentative and persuasive writing were salient in the rubrics' content feature that assessed these text types.

The subcategories generated within the organization category consisted of integrating the textual parts (introduction, body, conclusion, title), integrity, connection, and the train of thought. The subcategory with the highest frequency was the textual parts (64), which means organizing the parts of the text. The most common code was topical integrity (18). The rubrics emphasized the textual integrity by establishing connection among paragraphs/ideas/events/ sentences/ in the text flow, taking into consideration textual parts (introduction, body, conclusion), and train of thought.

No subcategories were generated from the category of style. Although most of the features were not, in fact, extricable, technically it did not seem possible to merge them altogether. The most common code was fluency (10). When we examined the rubric descriptions, we found that while the rubrics often highlighted the style/narrative features in writing, some rubrics - especially the adapted ones - referred to the writer. This should be considered natural in scoring the abovementioned feature. However, since the Turkish scientific tradition generally focuses on the stylistic feature of writing rather than the writer's style, fluency and originality were foregrounded in scoring, whereas criteria such as mood and risk-taking were not taken as equally significant. Because most of the features in this category bore subjectivity. The stylistic features varied widely based on the purpose of assessment but could not be merged into one category. The salient point was that while the rubrics did not include

stylistic dimensions as a category, the number of features related to style should not be underestimated.

A subcategory in presentation included line spacing for text units (24). Other codes could not be merged. The most common code was legibility (13). Considering the quality and quantity of the codes/features in this category, we can argue that the importance of the formal structure of writing (margins, line spacing, word spacing, line quality) was overemphasized.

Mechanics included compliance with spelling, punctuation and grammar rules, language use, and standard language use. The most frequent code was spelling (20). Apart from grammar, spelling, and punctuation, which plays a vital role in understanding the text, compliance with the rules, which the curriculum regards as essential, also stood out as a criterion for scoring.

No subcategory occurred in the vocabulary category. The most common code was word senses/word selection (19). The descriptions typically focused on selecting word senses and words in use accurately. However, although this category was about usage, it included a relatively small number of the concerning codes/features (note that it was mostly due to the difficulties in the descriptions with respect to the word feature).

In the sentence category, two categories were formed in terms of diversity and structure, and other codes did not generate subcategories. The most common codes were sentence meaning and different sentence structures (6). Connection between sentences consisted of many features. Following the coding procedure in this study, connection between sentences were included in organizational skills.

The textuality category with a small number of codes included cohesion and coherence (6). Coherence and cohesion are closely associated. Since most of the rubrics foregrounded textual cohesion, language and reasoning did not stand out. This has been an acceptable situation in the Turkish education system.

Process was the category with the smallest number of codes (5). The low frequency of codes in this category may also be considered reasonable since the rubrics generally scored complete writings.

Similarity ratio in descriptions

During the coding process, we aimed to determine the common or similar features in the descriptions, but since there were nuances within the similarities, it was difficult to present those findings within the scope of this study. Therefore, several important points will be highlighted here instead: According to the similarity matrix for codes calculated by the MAXQDA software, the similarity ratio ranged between 0.45 and 0.95. The similarity ratio between D3, which was at the center of the least similar rubric, and D10 and D22 was calculated as 0.45. The similarity ratio between D16a, which was at the center of the most similar rubrics, and D16b, D4, and D5 were 0,95. This similarity ratio was calculated, taking into account the number of similar codes. The ratio did not account for a 0.95 similarity or an almost complete similarity in terms of quality. For example, the number of similar codes between D4 and D16a was 15 (main idea, supporting ideas, topical integrity, connection among paragraphs, title, introduction, body, conclusion, spelling, punctuation, indenting, margins, line spacing, legibility, and word repetition). The number of similar codes between D3 and D10 was six (content-appropriateness for text type, title, conclusion, spelling, punctuation, word selection). The category to which the code belonged also affected the similarity ratio. Following this brief explanation, the similarity of the codes in the studies examined was displayed in Figure 1 upon analyzing the features as an entire set via MAXQDA software.

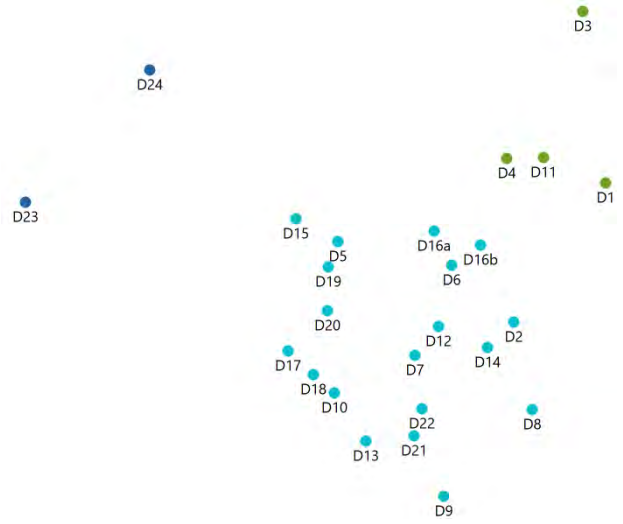

Figure 1. Similarity Map of The Features in The Rubrics

As shown in Figure 1, D1, D3, D4, D11, and D23, and D24 remained outside the blue cluster. One of the crucial reasons for the dissimilarity was the relatively large number of codes in the rubrics outside the cluster (see Table 3). Many criteria descriptions and orders in D23 and D24 were, in fact, not similar to the rest concerning the use of technical terms. The reason may be that these rubrics were adapted or translated versions. Since D1, D3, D4, and D11 showed less similarity in code relationships, they created a separate cluster. When the codes of these four rubrics were analyzed, there was a moderate similarity with the codes of the vocabulary category, which had a relatively small number of codes in the code pool, of the sentence category, and spacing as the subcategory of presentation. In other words, the relatively substantial similarity in only three categories distinguished these four rubrics from the rest. In the blue cluster, the similarity ratio was relatively higher due to the similarity in the features of spelling, punctuation, introduction, conclusion, topical integrity, and connection among paragraphs and the codes of the presentation category.

Based on Table 4 and Figure 1, we can argue that although the coding logic and the similarities in the descriptions of rubrics were an important factor in obtaining the findings, numerical findings were insufficient to explain the entire network of relationships. Although Figure 1 revealed the relationships and similarity ratios, the similarities appeared to vary considerably when we looked closely at the ratios individually.

Errors in The Descriptions of The Rubrics

In this review, various errors were identified in the criteria descriptions in the rubrics. We think that these errors were not related to the limitations that the rubrics inherited. These are considered essential to mention for those who will develop and use rubrics.

An example of an ambiguous description: "*Sentence length and style are expressive*" (D17). *Sentence style* is not a term or concept in the teaching of Turkish writing. It was ambiguous whether *style* meant sentence structure, style of expression, or syntax. Also, as to the Turkish grammar rules, this description might mean "the style of sentence length," which was again an ambiguous concept. In D18, the rubric used in the assessment of 4th-grade students' writing created ambiguity in meaning with the statement, "*The writer successfully uses words with new meanings*." According to this description, does the writer attribute a completely different meaning to a word or use an existing word with a new sense attributed by society? Or does the student attribute "more than one" new meaning to a word? The description was ambiguous in this respect.

An example of unnecessarily lengthy description*:* The rubric in D2 included the following descriptions to identify poor writing in terms of topical integrity (1); "*The writing has gone off-topic. The writing has shifted from the given topic. There are more than two statements in the text that shifted from the given topic.*" The first two descriptions had the same meaning. The third description, on the other hand, offered a particular condition for going off-topic. In other words, the three descriptions actually suggested that *there was no topical integrity*.

An example of a description that may cause inaccurate scoring: "*The writer changed the spelling to create style and used the spelling differently*" (D23). In fact, this description marked the use of spelling in a creative way. According to this rubric, a Turkish rater who reads the description of the excellent level for spelling criterion would rate *poor* (1) rather than *excellent* (5). The Turkish education system is prescriptive about spelling, and if you change the spelling rules or use it differently, you use the spelling incorrectly. For the poor level (1), the description of the spelling criteria "*The reader does not understand the text and has to reread it*" was ambiguous. It indicated that the incorrect use of spelling caused the reader/rater a comprehension problem. The problem arose due to different uses of spelling (it should also be noted that the rubric assessed the 5th-grade student writing).

An example case where the use of numbers complicates the assessment: "*At least seven of the eight word types -noun, verb, pronoun, adjective, adverb, postposition, conjunction, interjection- are used effectively*" (D14). Using this criterion for scoring required scrutinizing each sentence and identifying and counting the types of words. The job of the rater operating with such descriptions was quite challenging.

***An example of using the same feature in multiple criterion descriptions*:** In D19, the descriptions of two different criteria in the same dimension for the excellent level (3) were as follows: "*Writing sentences with no ambiguity*." "*Writing sentences that do not violate the subject-verb agreement, the object-verb agreement, and the indirect object-verb agreement*." The error indicated by the second description would already result in the ambiguity indicated in the first description. A rater had to score the sentence quality in writing twice for the same feature in such a case.

DISCUSSIONS

This study attempted to determine the procedures followed for validity and reliability of the writing rubrics in doctoral dissertations in Turkey, the dimensions constructing the rubrics, and the features assessed in writings through these rubrics. The findings indicated various deficiencies in the rubrics regarding validity and reliability and that the overlaps in dimension naming was reflected on the features. The analysis of the descriptions of 25 rubrics revealed nine categories and 104 features.

According to the findings, 88% of the writing rubrics did not conduct a validity analysis, while 20% had no reliability analysis. It appeared that reliability was regarded as relatively more important than validity. The literature supported the conclusion we arrived in the present study. Reddy and Andrade (2010) found that researchers attached more importance and focused more on reliability than validity. Humphry and Heldsinger (2014), pointing to the halo effect, stated that some studies failed to achieve validity. Therefore, considering the high number of studies without the validity of rubrics, we could argue that the expert opinion was insufficient to formulate an opinion. The fact that a significant part of the experts consulted were relatively inexperienced teachers may be another evidence of the inadequacy. Huot (2002), referring to many validity issues, argued that statistical calculations or operationalization of opinions to reach a decision were not a correct approach. Therefore, it is essential to provide both statistical and qualitative data about the validity of rubrics. For the validity of a rubric, the meticulous use of mixed methods, provided that they are coherent and complement each other, can eliminate some inadequacies.

Among the studies that provided the reliability analysis, two of them did not follow an appropriate method for the research (Miles-Huberman). The remaining 18 studies presented the inter-rater correlation coefficients to establish reliability. The studies reviewed in the analysis demonstrated that the inter-rater agreement was satisfactory or even better than satisfactory level. However, although we cannot argue which reliability analysis is sounder, we can point to which one is incomplete and insufficient. However, the rubric quality, the relatively low number of writings assessed, and insufficient information about the assessment process may cause some studies to be viewed with suspicion. Problematic issues regarding the reliability analysis of the rubrics included the lack of detailed evidence about the assessment process (Hodges et al., 2019), the presentation of mere correlation analysis (Wind, 2020), and incomplete or obscured information about the raters and inter-rater agreement. Previous studies (Behizadeh, & Engelhard, 2011; Eckes, 2008; Graham, Hebert, & Harris, 2015; Hodges et al., 2019; Humpry, & Heldsinger, 2014; Reddy, & Andrade, 2010) acknowledged significant deficiencies and inconsistencies in the assessment of writing regarding inter-rater agreement-hence reliability- for several reasons. In fact, Hodges et al. (2019) asserted that the presentation of inter-rater correlation coefficients alone should not be regarded as evidence for the reliability of a study. Following this claim, which even makes inter-rater agreement analysis controversial, we observed that five of the studies examined conducted no validity and reliability analyses, trivializing the rubric-based assessments. Researchers having conducted doctoral studies that lacked validity and reliability analyses, elaboration, rater training, and rating sessions and included dubious propositions need to focus more on these issues. Regarding rater training, Hodges et al. (2019) suggested that short-term training and a small number of rating sessions were inadequate. Similarly, Reddy and Andrade (2010) remarked that validity and reliability in inter-rater agreement should be taken more seriously and that there was a need for the careful use of accurate and functional research methods. Wind (2020) argued that the relationship among writing, rater, and rubric criteria could be uncovered with Many-Facet Rasch (MFR) models to reveal the details about the rating scale functioning.

Our findings revealed that the rubrics most commonly described three and five performance levels. Brookhart (2018), studying 51 rubrics, found that four and five were the most common performance levels. The differences in levels are acceptable, depending on the purpose of assessment as long as the assessment is conducted accurately, and the scale is divided into consistent levels. The findings also revealed that four (16%) of the twenty-five rubrics provided general descriptions of the criteria, not grading descriptions. Similarly, Brookhart (2018) emphasized that the performance levels of 14% of the rubrics she examined considered the criteria separately.

The codes obtained from the content analysis were categorized as follows: (genre-based) content, organization, style, presentation, mechanics, vocabulary, sentences, textuality, process. However, the number and frequency of features in textuality and process categories were quite low. Our attempt to classify the features revealed similar categories to the 6+1 Trait Writing Model (ideas, organization, voice, word choice, sentence fluency, convention, presentation). That is, content, style, mechanics, vocabulary, and sentence in our study overlapped with ideas, voice, convention, word choice, and sentence fluency, respectively. The organization and presentation categories had the same names for the features. Nevertheless, there was a clear difference between the descriptions of the rubrics examined and the ones in the 6+1 scoring rubrics. The descriptions in our analysis were simpler and more concise. On the other hand, the descriptions that put the writer or rater to the fore were extremely few. This may indicate more objective descriptions. Besides, the names of the categories we reached were similar, though not identical, to the categories (vocabulary, grammar, fluency, mechanics, coherence, style, topic) in Knoch's (2009) taxonomy. These similarities were actually due to the nature of writing. Despite the differences in assessment models, common to all models were textual features of grammar, vocabulary, and syntax (Knoch 2011). In large-scale writing assessments, the criteria typically include the train of thought, structure, completeness, description, argumentation, syntax, vocabulary, and correctness (Eckes, 2008). Therefore, we can argue that the features in these categories also overlapped with the commonly used assessment criteria.

As a result of the coding, we observed that the features regarding content and organization were frequent in the rubrics. From the first research conducted using writing rubrics until today, the cumulating evidence supported this finding. The research in this field highlighted that the raters often focused on content and organization (Huot, 1990; Huot, 2002). The frequency of features in the content and organization categories was higher than in the other seven categories. It showed that the descriptions in the rubrics used in the doctoral dissertations focused on these two categories. The features in the content category mostly focused on the technical processing of contextual information that the genre demanded, and the organization category allowed a holistic approach to text and to establish strong semantic and formalist links among the parts of a text.

The features in the style category were found to be quite varied. This variation in question is meaningful since the expected stylistic characteristics of writing differ based on the researchers' objectives and the writing genre. We asserted that the frequency of features in the presentation category was high, which suggested that the rubric makers focused too much on the outline and textual style. Considering that it was mostly the secondary school students' writings that were assessed, we observed that the formalist perspective and prescriptivism in the curriculum were projected in the rubrics. The previous curriculum (applied from 2006 to 2015) aimed to teach students to "pay attention to the outline and page layout, write properly and legibly, and learn cursive handwriting" (MEB, 2006, p. 29). The high number of features in mechanics can also be attributed to over-teaching of spelling and punctuation in the curriculum. Besides, many studies investigating the use of spelling and punctuation by secondary school students (e.g., Arı, & Keray, 2012; Bağcı, 2011; Karagül, 2010; Özbay, 1995; Süğümlü, 2020) stressed that spelling and punctuation were problematic for students, which may result in rubric makers concentrating on the descriptions of this feature.

The similarity matrices and the map in this study demonstrated that the number of features scored in rubrics, the complexity of rubrics, the extent (narrow/broad) of the descriptions of the features exhibited the differences among rubrics. In other words, the similarity decreased when the descriptions of the features overlapped. In this review, we found that most rubrics had short descriptions and that only one feature was prominent in rubrics, which marked the similarity of the rubrics to one another. A cursory purpose statement by researchers, text types indirectly measured by rubrics, and also the extent of descriptions in criteria may be the factors that increased or reduced the similarity of the features that rubrics aimed to measure.

The rubric examination showed the use of technical and academic terms and exhaustive descriptions and ambiguities in the descriptions in some rubrics. The language used in rubrics (Reddy, & Andrade, 2010), namely technical terms and adjectives that partially determine the level in quality definitions, can lead to ambiguity. Ambiguities in the rubric descriptions are frequently criticized (e.g., Chan, & Ho, 2019; Gottliebd, & Moroye, 2016). They can cause teachers or raters to have difficulty in assessing writing even if they understand the descriptions of features (Nordrum, Evans, & Gustafson, 2013).

As a result of the examination, we can claim that ambiguous descriptions, manipulations, quantifications, and the description of the same feature in two criteria with different statements can make the raters' job difficult. Some descriptions may require quantification. This was actually more desirable. There were many moderate-level descriptions (3) of the spelling criteria in the examined rubrics, such as "*There are between 2 and 5 errors*." With these descriptions, the rater can make rapid scoring decisions. However, if the excellent level of the criterion for word diversity was defined as "*using at least seven of the eight word types effectively*", the objective of student writing, the topic, and the differences in perspective would be ignored. Students may not use interjections and postpositions in the natural flow of writing. The rater may decide that five uses are effective, and two uses are not. The student may use seven types, two of which are not effective uses. In that case, it would be necessary to use numbers in the other performance descriptions. Such descriptions may lead to unfair treatment to the student who wrote the text and prevent the rater from making an informed decision.

The descriptions such as "*there are three..." or "two are used…*" should be rephrased, considering that they may result in an inaccurate measurement of learning (Brookhart, 2018).

CONCLUSIONS

This unique research suggested that the rubrics used in the doctoral dissertations in Turkey, like their universal counterparts, could list and describe the quality definitions of student writing in accordance with the objective of assessment, the scope, and the nature of writing. In recent years, it was observed that the researchers in Turkey were brave to prepare rubrics, and assessment with rubrics in doctoral research has increased. However, it should be noted that lacking validity and reliability analyses and inattentive descriptions can damage the measurements and the results of such important research as a doctoral dissertation. On the other hand, a small number of rubrics had problems regarding the use of wrong models and terms for the Turkish education system, ambiguity in descriptions, and overlapping features in two different criteria. For these reasons, the study of assessment where rubrics are used to assess writing has been in its infancy for about 20 years in Turkey. The research on computer-assisted assessment applications for student writing in some countries (e.g., Mao et al., 2018; Rahimi et al., 2017) indicates that Turkey needs to move forward with writing assessment research.

Our study presented that the rubrics focused on assessing secondary school students' writings in particular, and so far, not a single study has been conducted on the assessment of high school students' writings with rubrics. There are many reasons for the lack of research in this regard, but ultimately, we must realize that there is a crucial deficiency. A recommendation for short-term research is to assess the writings of high school students with rubrics. Another important recommendation is for policymakers, education administrators, educators, and researchers of writing to move along rapidly the infancy stage and decide on standardized assessment criteria conducting research on large-scale assessment. The categories and features proposed in this study can lay the groundwork. We suggest that researchers pay more attention to rater training to obtain more accurate and robust reliability, conduct analyses that corroborate validity and reliability, and present them in a transparent and detailed manner.

REFERENCES

Arı, G. (2008). *Öğrencilerin hikâye edici metinlerinin çözümleyici puanlama yönergesine göre değerlendirilmesi (6. ve 7. sınıf örneği)*. İstanbul: Marmara Üniversitesi yayımlanmamış doktora tezi

Arı, G. (2015). Yaratıcı yazma. *Üstün zekalı ve üstün yetenekli öğrencilerin eğitimi*. (ed. F. Şahin), pp. 127-172. Ankara: Pegem Akademi.

Arı, G., & Keray, B. (2012). Sekizinci sınıf öğrencilerinin noktalama işaretlerini uygulama düzeyi. *Electronic Journal of Social Sciences*, *11*(42), 40-54.

Babin, E., & Harrison K. (1999). *Contemporary composition studies a guide to theorist & terms*. Portsmouth: Grenwood Publishing.

Bağcı, H. (2011). Sekizinci sınıf öğrencilerinin noktalama işaretleri ile yazım kurallarını uygulayabilme düzeyi. *Turkish Studies, 6*(1), 672-684.

Behizadeh, N., & Engelhard Jr, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, *16*(3), 189-211.

Beyreli, L., & Arı, G. (2009). The use of analytic rubric in the assessment of writing performance-ınter-rater concordance study. *Educational Sciences: Theory and Practice, 9*(1), 105-125.

Brookhart, S. M. (2018). Appropriate criteria: key to effective rubrics. *In Frontiers in Education* (Vol. 3, p. 22). Frontiers. https://doi.org/10.3389/feduc.2018.00022

Brualdi, A. (2002). Implementing performance assessment in the classroom. İn Understanding scoring rubrics a guide for teachers, ed. C. Boston, (pp.1-4). Washington: Office of Educational Research and Improvement.

Chan, Z., & Ho, S. (2019). Good and bad practices in rubrics: The perspectives of students and educators. *Assessment & Evaluation in Higher Education, 44*(4), 533-545. https://doi.org/10.1080/02602938.2018.1522528

Coşkun, E. (2005). *İlköğretim öğrencilerinin öyküleyici anlatımlarında bağdaşıklık, tutarlılık ve metin elementleri*. Ankara: Gazi Üniversitesi Eğitim Bilimleri Enstitüsü yayımlanmamış doktora tezi.

Creehan, K. (1997). A discussion of analytic scoring for writing performance assessment. *In Annual Meeting of the Arizona Educational Research Association*, Phoenix, (ERIC Document Reproduction Service No. ED 414336).

Çakın, N. (1966). Yazma yolu üzerine. *Türk Dili*, 175, 489-490.

Deniz, K. (2000). *Yazılı anlatım becerileri bakımından köy ve kent beşinci sınıf öğrencilerinin durumu*. Çanakkale: Çanakkale Onsekiz Mart Üniversitesi Sosyal Bilimler Enstitüsü yayımlanmamış yüksek lisans tezi.

Duman, A. (1997). *Üniversitelerin Türk dili ve edebiyatı bölümleri dışındaki bölümlerinde Türkçenin eğitimi ve öğretimi*. Ankara: Gazi Üniversitesi Sosyal Bilimler Enstitüsü yayımlanmış doktora tezi.

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

Elbow, P. (2000). *Everyone can write essays toward a hopeful theory of writing and teaching writing*. New York: Oxford Universiy Press.

Enginarlar, H. (1990). *A contrastive analysis of writing in Turkish and English of Turkish high school students*. Ankara: Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü yayımlanmamış doktora tezi.

Erdiken, B. (1989). *Eskişehir sağırlar okulu ve Anadolu Üniversitesi İÇEM'de ortaokul sınıflarına devam eden 13-14 yaş işitme engelli öğrencilerin yazılı anlatım becerilerinin betimlenmesi*. Eskişehir: Anadolu Üniversitesi Sosyal Bilimler Enstitüsü yayımlanmamış yüksek lisans tezi

Erdiken, B. (1996). *Anadolu Üniversitesi İÇEM lise düzeyindeki işitme engelli öğrencilerin yazılı anlatım becerilerinin geliştirilmesinde işbirliği-gözlem yöntemi ile anlatım yönteminin karşılaştırılması*. Eskişehir: Anadolu Üniversitesi Sosyal Bilimler Enstitüsü yayımlanmamış doktora tezi.

Faigley, L., Cherry, R. D., Jolliffe, D., & Skinner, A. (1985). *Assessing writers: Knowledge and processes of composing*. Norwood: Ablex.

Flynn Jr, J. E., Tenam-Zemach, M., & Burns, L. D. (2015). Introduction: Why a book on rubrics? Problematizing the unquestioned. *Rubric nation: Critical inquiries on the impacts of rubrics in education*, pp. xi-xxx. Charlotte: Information Age Publishing.

Glass, K. T. (2005). Curriculum design for writing instruction. California: Corwin Press.

Gottlieb, D., & Moroye, C. M. (2016). The perceptive imperative: Connoisseurship and the temptation of rubrics. *Journal of Curriculum and Pedagogy*, *13*(2), 104-120. https://doi.org/10.1080/15505170.2016.1191389

Göğüş, B. (1971). Ana dili olarak Türkçenin öğretimine tarihsel bir bakış. *TDAY-Belleten 1970*, 123-154.

Göğüş, B. (1978). *Orta dereceli okullarımızda Türkçe ve yazın eğitimi*. Ankara: Gül

Göğüş, B., Yücesan, S. (1988). *Uluslararası Anadili Eğitim Örgütü Çalışmaları çerçevesinde Türkiye'de bir Türkçe eğitimi portresi*. Ankara: Rehber.

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

Graham, S., Harris, K., and Hebert, M. A. (2011). *Informing writing: The benefits of formative assessment. A Carnegie Corporation Time to Act report*. Washington, DC: Alliance for Excellent Education.

Graham, S., Hebert, M. A., Harris, K. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal, 115*(4), 523–547. https://doi.org/10.1086/681947

Gunning, T. G. (2006). *Assessing and correcting reading and writing difficulties*. Boston: Pearson Education Inc., 3. edition

Hodges, T. S., Wright, K. L., Wind, S. A., Matthews, S. D., Zimmer, W. K., & McTigue, E. (2019). Developing and examining validity evidence for the writing rubric to ınform teacher educators (WRITE). *Assessing Writing*, 40, 1-13. https://doi.org/10.1016/j.asw.2019.03.001

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher*, *43*(5), 253–263. doi:10.3102/0013189x14542154

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*(2), 237-263.

Huot, B. (2002). *Rearticulating writing assessment for teaching and learning*. Logan: All USU Press Publications. (https://digitalcommons.usu.edu/usupress_pubs/137)

Jokobson, J. R. (2005). Six traits writing using literature as a model. *Books Links*, *14*(5), 44-47.

Karagül, S. (2010). *İlköğretim 6-8. sınıf öğrencilerinin Türkçe Dersi Öğretim Programı'nda belirtilen yazım ve noktalama kurallarını uygulayabilme düzeyi*. İzmir: Dokuz Eylül Üniversitesi yayımlanmamış yüksek lisans tezi.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from?. *Assessing Writing*, *16*(2), 81-96.

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

Levend, A. S. (1954). Ana dilini doğru yazma geleneği bizde niçin yoktur? *Türk Dili*, 29, 251-253.

Li, J., & Lindsey, P. (2015). Understanding variations between student and teacher application of rubrics. *Assessing Writing*, 26, 67-79. doi:10.1016/j.asw.2015.07.003

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H. S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment, 23*(2), 121-138. https://doi.org/10.1080/10627197.2018.1427570

Martin-Kniep, G. O. (2000). *Becoming a beter teacher: Eight innovations that work*. Alexandria: Association for Supervision & Curriculum Development.

MEB (2006). *İlköğretim Türkçe Dersi (6, 7, 8. Sınıflar) Öğretim Programı*. Ankara: MEB.

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

Nordrum, L., Evans, K., & Gustafsson, M. (2013). Comparing student learning experiences of in-text commentary and rubric-articulated feedback: strategies for formative assessment. *Assessment & Evaluation in Higher Education*, *38*(8), 919-940. DOI:10.1080/02602938.2012.758229

O'Neill, P., Moore, C., & Huot, B.(2009). *A guide to college writing assessment*. Logan: Utah State University Press.

Özbay, M. (1995). *Ankara merkez ortaokullarındaki üçüncü sınıf öğrencilerinin yazılı anlatım becerileri üzerine bir araştırma*. Ankara: Gazi Üniversitesi Sosyal Bilimler Enstitüsü yayımlanmış doktora tezi.

Özdemir, E. (1983). Ana dili öğretimi. *Türk Dili*, 379/380, 18-30.

Pehlivan, A. (1994). *Kuzey Kıbrıs Türk Cumhuriyeti ilkokul öğrencilerinin yazılı anlatımlarında görülen özellikler*. Ankara: Ankara Üniversitesi Sosyal Bilimler Enstitüsü yayımlanmamış yüksek lisans tezi.

Popham, W. J. (1997). What's wrong and what's right with rubrics? *Educational Leadership*, *55*(2), 72-75.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders*. Boston: Allyn & Bacon.

Rahimi, Z., Litman, D., Correnti, R., Wang, E., & Matsumura, L. C. (2017). Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education, 27*(4), 694-728. https://doi.org/10.1007/s40593-017-0143-2

Rankin, A. D. (2015). *A comparability study on differences between scores of handwritten and typed responses on a large-scale writing assessment*. The University of Iowa unpublished doctoral dissertation

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

Sarıca, H. Ç., & Usluel, Y. K. (2016). Eğitsel bağlamda dijital hikâye anlatımı: bir rubrik geliştirme çalışması. *Eğitim Teknolojisi Kuram ve Uygulama, 6*(2), 65-84. https://doi.org/10.17943/etku.12600

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

Sever, S. (1993). *Türkçe öğretiminde uygulanan tam öğrenme kuramı ilkelerinin, öğrencilerin okuduğunu anlama ve yazılı anlatım becerilerindeki erişiye etkisi*. Ankara: Ankara Üniversitesi Sosyal Bilimler Enstitüsü yayımlanmamış doktora tezi.

Sezer, S. (2005). Öğrencilerin akademik başarısının belirlenmesinde tamamlayıcı değerlendirme aracı olarak rubrik kullanımı üzerinde bir araştırma. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 18, 72-84.

Shermis, M. D., Lottridge, S., & Mayfield, E. (2015). The impact of anonymization for automated essay scoring. *Journal of Educational Measurement, 52*(4), 419-436. https://doi.org/10.1111/jedm.12093

Spandel, V. (1997). *A handbook for parents of 6 trait writing students*. Portland: Northwest Regional Educational Laboratory.

Süğümlü, Ü. (2020). *Ortaokul öğrencilerinin yazma çalışmalarındaki yazım ve noktalama hatalarının belirlenmesi*. Ana Dili Eğitimi Dergisi, *8*(2), 528-542. doi.org/10.16916/aded.706748

Şengiz, S. (1976). Öğrenci kompozisyonları. *Türk Dili*, 303, 745-747.

Şimşek, Ö. (2000). *İlköğretim 4. ve 5. sınıf öğrencilerinin yazılı anlatım becerilerinin incelenmesi ve bilgisayar destekli yazılı anlatım*. İstanbul: Marmara Üniversitesi Eğitim Bilimleri Enstitüsü yayımlanmamış yüksek lisans tezi.

Şimşek, R. (1983). Çağdaş eğitimde anadilinin yeri. *Türk Dili*, 379/380, 36-39.

Tekşan, K. (2001): *Yazılı anlatımı geliştirmede ön hazırlığın etkisi*. Çanakkale: Çanakkale Onsekiz Mart Üniversitesi Sosyal Bilimler Enstitüsü yayımlanmamış doktora tezi.

Temur, T. (2001). *İlköğretim 5. sınıf öğrencilerinin yazılı anlatım becerilerinin beceri düzeyleri ile okul başarıları arasındaki ilişki*. Ankara üniversitesi sosyal bilimler enstitüsü yayımlanmamış yüksek lisans tezi.

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

White, E. M. (1985). *Teaching and assessing: Recent advances in understanding evaluating and improving student performance*. San Francisco: Jossey-Bass.

Wilson, M. (2006). *Rethinking rubrics in writing assessment*. Portsmouth: Heinemann

Wind, S. A. (2020). Do raters use rating scale categories consistently across analytic rubric domains in writing assessment?. *Assessing Writing*, 43, 100416. https://doi.org/10.1016/j.asw.2019.100416