Croatian Journal of Education
Vol.21; No.1/2019, pages: 103-155
Original research paper
Paper submitted: 12th September 2017
Paper accepted: 4th September 2018
https://doi.org/10.15516/cje.v21i1.2922

Evaluating Essay Assessment: Teacher-Developed Criteria versus Rubrics. Intra/Inter Reliability and Teachers' Opinions

Veda Aslim-Yetis
Anadolu University, Faculty of Education

Abstract

Rater reliability plays a key role in essay assessment, which has to be valid, reliable and effective. The aims of this study are: to determine intra/inter reliability variations based on two sets of grades that five teachers/raters produced while assessing argumentative essays written by 10 students learning French as a foreign language in accordance with the criteria they had developed and with a rubric; to understand the criteria they used in the assessment process; and to note what the raters/teachers who used rubrics for the first time within the scope of this study think about rubrics. Quantitative data set has revealed that intra-rater reliability between the grades assigned, through the use of teacher-developed criteria and the rubrics, is low, that inter-rater reliability is again low for the grades based on teacher-developed criteria, and that inter-rater reliability is more consistent for assessments completed through the use of rubrics. Qualitative data obtained during individual interviews have shown that raters employed different criteria. During the second round of individual interviews following the use of rubrics, raters have noted that rubrics helped them to become more objective, contributed positively to the assessment process, and can be utilized to support students' learning and to enhance teachers' instruction.

Key words: *evaluation; mixed-method research design; writing.*

Introduction

Foreign language instruction is directed toward improving four basic skills (reading, writing, speaking, and listening) and their relevant sub-skills. Writing, as one of

the basic skills, requires the use of many sub-skills in order to convey a message successfully. This skill does not only require the correct use of linguistic knowledge (grammar, vocabulary, spelling, etc.) but also entails production of genuine ideas, organization of those ideas in a consistent layout, clear expression of thoughts, creating interest in readers, and being comprehensible. In other words, students have to have control over their texts in terms of content, style, organization of ideas, text type (descriptive, argumentative, etc.), linguistic rules of the target language, and the conventions (punctuation, upper/lower case use, paragraphing, etc.). Likewise, teachers have to examine carefully all the sub-skills embedded in writing and sort out significant amount of information while "making judgments about quality – how good the behavior or performance is" (McMillan, 2004, p. 10). Therefore, assessing written works appears to be laborious, weary, and time-consuming. That is why, "For many years, […], writing assessment has been plagued by concerns about the reliability of rating (which usually means, the reliability of raters)" (Hamp-Lyons, 2007, p. 1).

> Reliability refers to the consistency of assessment scores. For example, on a reliable test, a student would expect to attain the same score regardless of when the student completed the assessment, when the response was scored, and who scored the response (Moskal & Leydens, 2000, n.p.).

Assessing written works objectively, or at least as objectively as possible, matters significantly. The probability that judgments concerning the correctness of long-winded answers may vary across raters jeopardizes the reliability and validity of assessment (Ozcelik, 1992, p. 127). Assessing essays using objective criteria lowers the rate of mistakes in the assessment process to a minimum and produces more impartial and correct results about students' skills. However, it is not feasible to eliminate the rate of mistakes totally from the assessment procedure. Since learning is an abstract process, and it is not possible to directly assess writing skills, forms of indirect assessment are employed to collect data concerning students' writing skills. Besides, more mistakes stemming from students, raters, assessment tool or method, and the setting can interfere with evaluation and assessment process. For instance, several factors, such as scoring the papers at different times, physiological (fatigue) and psychological (being joyful or not) state, and time of assessment (too early/late) can allow for mistakes in the assessment procedure. Even positive or negative feelings that a rater bears for his/her student may influence him/her and lead to mistaken assessments. Thus, assessment of a paper by the same rater at different times may not produce the same score, and students may have to accept drastically different grades although their works on the paper are fairly similar. The inconsistencies emerging from internal and external factors during the assessment process concern issues relating to intra-rater reliability (Moskal & Leydens, 2000).

Furthermore, different grades may be the result of employing different criteria or assigning different scores to the same criteria during the assessment of written works. For instance, grammatical and structural features matter more for some raters, whereas

language use and debate skills deserve higher scores for some other raters (Moskal, 2000). Still, some raters judge a student's performance based on the text s/he produces, and others evaluate the same performance through comparison with other students' texts and performances (Romainville, 2011). Therefore, the reliability of assessment procedure is reduced and drifts away from objectivity. These inconsistencies, stemming from raters' idiosyncratic and different assessment criteria, are a matter of inter-rater reliability (Moskal & Leydens, 2000).

What a rater or raters are expected to do is to perform reliable assessment producing similar or the same results free from personal judgments, factors outside instructional goals, and temporal and spatial limits. In order to achieve this, assessment criteria should be determined in advance, and all students' works should be scored in accordance with those criteria.

The analysis of relevant literature has yielded that scoring rubrics is one of the efficient tools to be employed in order to prevent raters from ignoring assessment criteria, changing (either consciously or sub-consciously) these criteria from essay to essay, and to display assessment criteria as a whole for performance assessments, such as writing skill (Berthiaume et al., 2011; Brookhart & Chen, 2015; East, 2009; Jonsson & Svingby, 2007; Moskal, 2000; Moskal & Leydens, 2000; Nitko, 2004; Reddy & Andrade, 2010; Scallon, 2004). Accordingly, Moskal (2000, p. 1) states: "Scoring rubrics are typically employed when a judgement of quality is required and may be used to evaluate a broad range of subjects and activities." Andrade (2005, p. 27) defines a rubric as follows:

> "Rubric is an assessment tool that lists the criteria for a piece of work or what counts (for example, purpose, organization, details, voice, and mechanics often are what count in a written essay) and articulates gradations of quality for each criterion, from excellent to poor."

Designed in accordance with certain criteria, and designating what to score for each criterion, scoring rubrics are gradient and descriptive scoring tools employed to sustain a standard and stable assessment. These tools are either in a ready-made form or can be developed by teachers/raters, based on the qualities of written works (Stevens & Levi, 2005). Rubrics are of two kinds: holistic and analytic. Holistic rubrics "rate or score the product or process as a whole without first scoring parts or components separately" (Nitko, 2004, p. 264). With a span between 3 to 6 points, holistic scoring entails grading the text globally by assessing the written work as a whole, thereby enabling a quick and easy assessment, which is often a factor of choice when working with crowded classes. This kind of assessment mostly focuses on the final product rather than the process that may have been influential over students' performance. For instance, a 3-scale holistic assessment rubric, designed for writing skills, is as follows: Excellent (3 points) – clear expression of thoughts, opinions are supported with examples, no spelling mistakes, no grammatical mistakes; Good (2 points) – thoughts are understandable, examples do not quite match the opinions, spelling is

mostly correct, few grammatical mistakes; Poor (1 point) – thoughts are not clear, opinions are not supported with examples, there are many spelling mistakes, quite a few grammatical mistakes. In accordance with this scoring, a student's work can be graded with a grade 1 to 3. However, these rubrics do not provide enough feedback in order for students to improve their performance and do not allow separate analysis of each criterion, because it handles all criteria holistically. Thus, studies concluding that analytic scoring rather than holistic scoring is more suitable and productive for assessing writing skills are becoming more and more common (Brookhart, 2013; Lumley, 2002).

"Analytic rubrics describe work on each criterion separately" (Brookhart, 2013, p. 6). "These rubrics rate or score separate parts or characteristics of the product or process first, then sum these part scores to obtain a total score" (Nitko, 2004, p. 264). Therefore, when used to assess writing skills, these rubrics rate each component of writing separately in a gradient manner, allowing detailed scoring for each part. For instance, these rubrics focus on several topics, such as task completion, content, expression of thoughts, word choice, and grammar, and assign a gradient scoring key for these topics. A text may be scored with grade 4 with respect to content, with 5 for word choice and grammar, and 2 in terms of syntax. Teachers better understand how to score each criterion, and students realize what criterion is poor for them and know that they have to consider and eliminate these weak points for their following writing task. In other words, these rubrics identify students' strong and weak sub-skills, document teachers' rationale for their assessment, and introduce the opportunity to provide more feedback on students' less-developed skills rather than simply writing a few vague words on students' papers. As stated by Jonsson and Svingby (2007, p. 132): "Analytical scoring is useful in the classroom since the results can help teachers and students identify students' strengths and learning needs."

The overall aim of this study is to determine the reliability of rubrics in assessing writing performance and to identify their effect over inter/intra-rater reliability. Accordingly, answers have been sought for the following research questions:

1. Does intra-rater reliability vary across writing scores given through teacher-developed criteria (TDC) and scoring rubrics?
2. Does inter-rater reliability vary across writing scores given through TDC and scoring rubrics?
3. What are the assessment criteria and scoring systems that participating teachers/raters employ?
4. What do the raters think about using rubrics?

Methodology

Research Model

As one of the mixed-method research designs where qualitative and quantitative research data are collected and analyzed, sequential explanatory design has been

employed for this study in order to be able to answer the research questions. In this design, research commences with the collection and analysis of quantitative data that would answer the research questions and continues with in-depth analysis of quantitative data through the use of qualitative data to be able to make interpretations (Creswell & Plano Clark, 2011).

In this research, quantitative data were collected via assessment of writing papers in accordance with both TDC and rubric scoring, and qualitative data were obtained from semi-structured individual interviews.

Participants

The research has been conducted with five instructors working at French Language Program of Foreign Languages Schools affiliated with two universities in Turkey. The instructors participated in the study voluntarily.

Three primary selection criteria were used in order to choose the participating instructors: at least two years of experience in teaching French language, teaching writing course, and no experience in using rubrics. As dictated by ethics, instructors' names are not provided in order to protect their privacy. General characteristics of teachers/raters (R) are as follows.

Rater 1: Aged 30, R1 has been teaching French at the Foreign Languages School for five years. R1 teaches "Reading" as well as "Writing". S/he also has experience teaching speaking classes. R1 has not worked at any other institution previously.

*Rater 2***:** Aged 27, R2 has been teaching French at the Foreign Languages School for two years. S/he teaches only "Writing" and has not worked at any other institution previously.

Rater 3: Aged 28, R3 has been teaching French at the Foreign Languages School for four years. S/he teaches a "Language Activities" course where miscellaneous activities are conducted along with writing. S/he has not worked at any other institution previously.

Rater 4: Aged 32, R4 has been teaching French at the Foreign Languages School for four years. S/he teaches "Reading", "Speaking", and "Listening" courses as well as "Writing". S/he has not worked at any other institution previously.

Rater 5: Aged 35, R5 has been teaching French at the Foreign Languages School for three years. S/he teaches "Reading" as well as "Writing". S/he has not worked at any other institution previously.

Data Collection Instruments/Tools

Essays

The essays that were assessed within the scope of this research belonged to 10 B1-level students learning French as a foreign language. Each student wrote one argumentative/persuasive essay for the final exam of their "Writing" class. The reason

why the researcher chose exam papers was the hope that students would develop their essays more seriously. All students were informed about the research, and their consent was granted prior to the application.

The instruction for the argumentative essay was as follows: "*Do you think it is a good idea to go abroad for education?* State your opinion and support it with valid arguments. Produce clear and coherent writing in which the development, organization, and style are appropriate to task, purpose, and audience. Time: 90 min. Words: 160-180." The rationale behind selecting an argumentative essay is the fact that it requires individuals/writers to produce ideas based on realistic reasons and components of reality and present them highly consistently. The writer has a major cognitive responsibility not to write and report unreal things but to note true and real information as much as possible. An argumentative (or persuasive) essay aims to convince readers, impose an idea, and encourage readers to do or not to do something. Thus, the writer has to search for and present reasonable and convincing arguments (Tompkins, 2004, p. 421).

Actually, 20 students took the exam, but the study was restricted to 10 students in order to minimize the rate of possible mistakes that would interfere with the assessment procedure. As a matter of fact, five raters noted that the number of papers was reasonable and feasible during the pre-interviews. The following is how 10 texts were chosen: a total of 20 essays written by 20 students who had been attending the writing class and who took the final exam were analyzed, and 10 of them were selected in terms of the number of ideas, consistent presentation of ideas, and length.

Semi-structured Individual Interviews during the First Step

The goal of semi-structured interviews held right after the raters graded the essays based on teacher-developed criteria was to determine essay assessment criteria that raters either considered or ignored, to depict personal theories and performance constructs that raters based their assessment on, and to understand their scoring system.

Accordingly, raters were asked open-ended questions to figure out which criteria they attended and which ones they ignored during assessment. Moreover, the raters were reminded of other criteria that they had not taken into account and were asked why they had not attended or considered them.

Analytic Rubric

This study employed an analytic rubric prepared (in French) for the assessment of argumentative essays. With a total score of 25 points, this rubric was developed by CIEP (International Center for Pedagogical Studies) in order to assess argumentative essays written by B1-level students (Breton et al., 2010, p. 74). The rubric has three parts: part one includes items about task completion, coherence-cohesion, and clear expression of ideas (a total of 13 points); part two is composed of items regarding

lexical competence/lexical spelling (a total of 6 points) and includes vocabulary range and control and orthographic control; and part three focuses on grammatical competence/grammatical spelling (a total of 6 points) and consists of items about sentence structure, tense and mood, and morpho-syntax.

In a nutshell, the first part evaluates if the components of argumentative essay outlined in the instruction can be identified or not (e.g. word count, presentation of arguments, etc.) and assesses the principles of writing skill whereas the second and third parts regard mostly grammar and vocabulary knowledge that helps increase the quality of a written work. As noted by Simard (1992, p. 286), the writer, unlike the speaker, has to be clearer, more understandable, more explicit to express his ideas by using "a wider and more diverse linguistic repertoire." Explanations concerning each item are provided within the rubric itself (see Appendix; English version), and to get a score out of 100, raters have to add up the total score obtained from the rubric, and simply multiply it by 4.

Semi-structured Individual Interviews during the Second Step

These semi-structured interviews were conducted immediately after the raters graded the set of 10 essays (which they had scored without the rubrics three months ago), based on the analytic rubric set. The goal of these interviews was to note raters' opinions about using rubrics, if it had been practical or not to use them, and the positive and negative aspects of using rubrics.

Practice/Implementation

Prior to the study, French language teachers working at a Foreign Languages School were contacted and informed about the research. Five teachers willing to participate and matching the selection criteria were chosen, and they were given detailed information about the study and their tasks, which included double assessment of 10 essays written by French language students. Meanwhile, they were also assured that no personal data would be disclosed, and none of their remarks would be associated with them.

Subsequently, each teacher graded the 10 essays on different days by using their own teacher-developed criteria. Essays were numbered randomly, and teachers were instructed to grade the essays in the same order. Since the grading system in Turkish universities is in the range 1-100, teachers were told to score the papers with numbers up to 100. Raters were left alone in a room in order to minimize the distractors and to guarantee no disturbance during grading. Semi-structured individual interviews were held right after the grading sessions.

An interval of three months was considered enough time in order for the raters not to remember the contents of the essays and their grades before starting the second step of the research. Three months later, raters were again invited on different days.

This time, the raters were asked to grade the same argumentative papers using an analytic rubric set. In addition, they were again reminded to follow the same order as outlined by the numbers on the papers, which helped in keeping the order of scoring the same for both with and without analytic rubric grading. Scoring was again done with numbers up to 100. Similar to the first step, raters were left alone in a room for their assessment, but a short session (20 min) was held with each rater to discuss the rubric set they would use and to clarify any unclear points about the rubrics before they started grading. Right after the grading sessions, semi-structured interview sessions were conducted with each rater.

Results

Does Intra-Rater Reliability Vary across Writing Scores Given through Teacher-Developed Criteria (TDC) and Scoring Rubrics?

Table 1 displays the difference between scores that raters assigned to each essay by using teacher-developed criteria (TDC) and rubrics, the number of essays within the same grade range, and information about intra-reliability of raters.

Table 1

Intra-rater consensus between TDC and rubric assessment

Differences	Rater 1 F*	Rater 2 F	Rater 3 F	Rater 4 F	Rater 5 F
0	0	0	0	0	0
±1-5	2	3	1	4	0
±6-10	1	3	6	2	5
±11-15	2	2	1	1	1
±16 and more	5	2	2	3	4
TOTAL	10	10	10	10	10

*Frequency

A closer look at Table 1 yields that none of the raters gave the same score to any of the essays they had graded earlier. Rater 1 (R1) holds the lowest level of consensus by giving scores ranging between ±16 to five of the 10 papers. Similarly, R5 bears the second lowest level of consensus, since s/he assigned scores ranging between ±16 to four of the essays s/he had graded. Whereas R1, R2, R3, and R4 assigned scores between ±1-5, R5 scored none of the essays within this narrow range and assigned scores between ±6-10 to five of the papers. Based on the results in this table, one can simply conclude that raters' scores display variation and that intra-reliability is considerably low. The correlational coefficient between assessments with and without rubrics was analyzed in order to better comprehend rater reliability of two assessments with and without rubrics and to be able to reach a statistical definition. In this regard, Spearman's rank correlation coefficient (R) – the non-parametric equivalent of Pearson's product moments correlation – was employed (Table 2).

Table 2
Correlations across TDC and Rubric Assessments

Raters	R
1	-.353
2	.631
3	.494
4	.433
5	-.016

As shown in Table 2, the correlation coefficients range between -.016 and .631, and none of these five values is statistically significant, which means that all raters gave different scores to the same 10 essays when they used rubrics rather than teacher-developed criteria. Therefore, it is possible to state that raters do not have a consistent scoring system, and that intra-rater reliability of the writing scores is pretty low. As a matter of fact, Table 3 points out that the difference between scores assigned by using TDS and the rubrics range from 2 up to 38. For instance, R3 gave grade 88 to the fourth essay that s/he had previously graded with a 50 (+38 points). Likewise, R5 gave a 92 to the eighth essay that s/he had previously graded with a 55 (+37). The same rater scored the seventh essay that s/he had given a 68 with a 90 (+22), and the score 90 that R5 found suitable for the third essay based on TDC was lowered to a 68 (-22) based on rubric scoring. Both small- and large-scale discrepancies are observed among all the scores given by all the raters.

Table 3
Scores based on TDC and analytic rubrics

Essays	Assessment via TDC					Assessment via Rubrics				
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
1	75	85	55	78	88	80	74	62	70	78
2	72	85	70	82	90	78	72	76	78	74
3	70	80	60	85	90	72	70	70	66	68
4	72	85	50	65	90	89	88	88	82	82
5	86	80	55	65	60	72	72	74	70	70
6	94	80	65	85	70	78	77	74	78	80
7	74	98	85	92	68	90	88	88	89	90
8	76	95	70	80	55	90	90	82	83	75
9	85	80	60	65	73	58	52	56	61	58
10	78	85	50	75	60	60	56	58	58	54

Does Inter-Rater Reliability Vary across Writing Scores Given through TDC and Scoring Rubrics?

Table 3 shows that not only the scores given by the same rater to the same essay change, depending on the assessment tool, but also the scores given by different raters to the same essay differ as well. For instance, scores given to the third, fifth, and seventh

essays range between 60-90, 55-86, and 68-98, respectively, which indicates that the differences between scores are a major problem, especially for assessments based on TDC. On the contrary, the scores given to the same essays by the same raters, based on the rubrics, do not range that drastically, and there is somewhat of a consensus between raters: scores for the third essay are 66 and 72, for the fifth essay 70 and 74, and for the seventh essay 88 and 90. The scores given by five raters based on TDC and the rubrics are shown in the following two graphs in order to depict this consensus more clearly (Figures 1 and 2).

Figure 1. Distribution of the scores based on TDC

Figure 1 shows the distribution of the scores assigned by each rater to 10 essays, based on TDC, and it is visible that each rater gave a different score, which indicates a lack of consensus and existence of strong divergence.

However, an examination of Figure 2, which displays the scores assigned via use of rubrics, reveals that distributions are closer, and even the same most of the time. As can be seen in this figure, curves symbolizing raters and their grades generally overlap, which means that assessments were similar, and a consensus was achieved by using rubrics. A possible interpretation of this finding is that assessing writing papers based on rubrics produces more consistent scores.

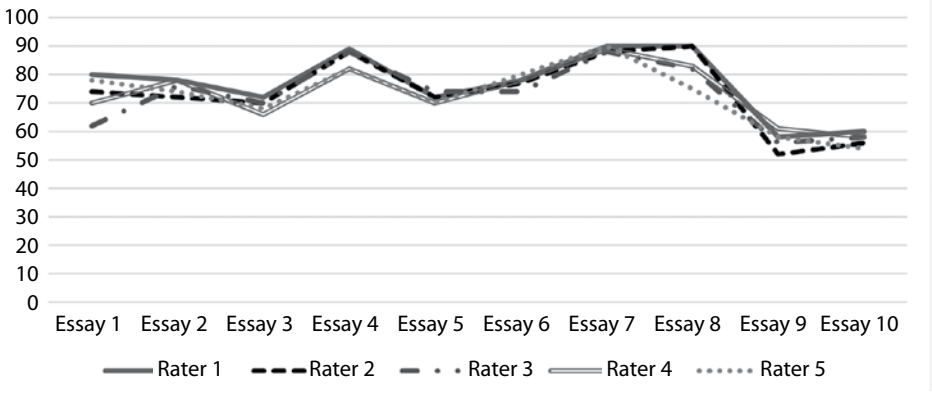

Figure 2. Distribution of the scores based on rubrics

Kendall's Coefficient of Concordance (Kendall's tau (W)) was employed in order to statistically determine scoring reliability of raters who used both TDC and rubrics. A non-parametric statistical analysis, Kendall's Coefficient of Concordance, is used to measure agreement between more than two raters who have their own judgments about qualitative categories and to determine reliability and compatibility among raters (Akbulut, 2010, p. 174; Alpar, 2012, p. 464). According to Kendall's tau (W), the coefficient is between 0 (no compatibility) and +1 (full compatibility), and this value points to a high level of compatibility if it is closer to 1 (Can, 2014, p. 376). Within the scope of this research, the compatibility among the raters was analyzed, first, for the scores given in accordance with TDC and, then, for the scores given in accordance with rubrics (Table 4).

Table 4

Kendall's tau (W) coefficient of concordance across scores given via TDC and via rubrics

Raters (N)		W	X^2	Df	p
5	TDC	.170	7.633	9	.572
	Rubrics	.903	40.654	9	.000

Relevant calculations completed with the values in Table 4 yielded that the Kendall's tau (W) coefficient was determined to be 0.17 ($p>0.05$) for the assessments based on TDC. Being so close to 0, this value indicates a very low level of compatibility among the scores given by five raters. However, the Kendall's tau (W) coefficient for the assessments based on rubrics was calculated as $W=0.903$ ($p<0.05$). Considering that it is very close to 1, it is possible to note that the compatibility among the scores given by five raters based on rubric scoring is quite high.

As a result, it is statistically proven that inter-rater reliability for writing assessments completed through the use of rubrics is higher than other grading procedures based on TDC.

What Are the Assessment Criteria and Scoring Systems That Participating Teachers/Raters Employ?

The findings obtained during semi-structured individual interviews held in the first step in accordance with the third research question are first presented in a table form (Table 5) and then discussed in detail beneath the tables.

Table 5 indicates that R1, R2, R3, R4, and R5 employed between four and seven criteria when they assessed the papers according to their own set of criteria. Yet, there are three criteria that all the raters used during their evaluation: text features, grammar, and vocabulary. Besides, there is another noteworthy difference with respect to the content of the criteria. For instance, R5 did not mention spelling, yet s/he split grammar into two (grammatical spelling and lexical spelling), and assessed spelling within lexical spelling criterion.

Table 5

Findings of the first semi-structured interviews

Raters	Assessment criteria raters considered	Assessment criteria raters ignored	Resulting scoring system
R1	* Text Features * Grammar * Vocabulary * Coherence/cohesion	* Task completion * Syntax * Spelling	* Text features: 50 points (p.) – Strength of arguments – Paragraphing – Are there an introduction, the main body and a conclusion? *Grammar: 15p. – (-1) point for each mistake (only 1 point is subtracted for recurring mistakes) * Vocabulary: 15p. – Overall assessment * Coherence/cohesion: 20p. – Overall assessment
R2	* Task completion * Grammar * Spelling * Vocabulary * Text Length * Text features	*Coherence/ cohesion	* Task completion: 10p. * Text length: 5p. * Text features: 15p. – Arguments – Thesis/anti-thesis For the remaining 70p.: * Grammar: (-½) points for each mistake * Spelling: (-½) points for each mistake * Vocabulary: (-½) points for each mistake
R3	* Coherence/cohesion * Vocabulary * Grammar * Text features	* Spelling * Task completion * Syntax	* The system is binary (either 1 or 0) 0 point for each mistake +1 points for each correct use * Vocabulary overall assessment (no use of binary system) * Overall assessment of arguments and text features (no use of binary system) * Scoring: – 100/4= 25p. – Grammar: 25p. – Coherence/cohesion: 25p. – Vocabulary: 25p. – Arguments/text features: 25p.
R4	* Task completion * Vocabulary * Grammar * Spelling * Coherence/cohesion * Expression of opinions * Text features	* Syntax	* Task completion: 5p. * Vocabulary: 10p. * Grammar: 15p. * Spelling: 15p. * Coherence/cohesion: 20p. * Expression of opinions: 20p. * Text features: 15p. –Text structure – Introduction? ➔ Overall assessment for all

Raters	Assessment criteria raters considered	Assessment criteria raters ignored	Resulting scoring system
R5	* Text features * Vocabulary * Grammar: grammatical spelling and lexical spelling *Coherence/cohesion	* Task completion * Syntax	* Scoring: – 100/4= 25p. – Text features: 25p. – Grammar: 25p. – Coherence/cohesion: 25p. – Vocabulary: 25p. ➔ Overall assessment for all

As again shown in Table 5, R1, R2, R3, R4, and R5 ignored between one and three criteria during the assessment of the essays. R1 and R3 neglected the same set of three criteria (task completion, syntax, spelling), R2 did not employ the criterion of coherence/cohesion, R4 ignored syntax, and R5 did not include task completion into the assessment procedure.

Concerning the neglected criteria, R1 said the following about task completion: "To me, this criterion is embedded within text features. I do not regard this as a separate criterion." As for Syntax, the same rater reported that it is the same thing with grammar, and thus evaluated within grammar. Regarding spelling, R1 stated that spelling is not a major concern and said: "What matters is to understand what the student means. If it is comprehensible, spelling mistakes are not that important", and added, "If it is not understandable, I consider that within coherence/cohesion, and subtract points if necessary." It seems impossible to agree with this rater, since spelling is a crucial component of a well-written essay, and poor spelling interferes significantly with the comprehensibility of the text by clouding the clarity of meaning and tiring the readers.

R2 explains why s/he did not take "coherence/cohesion" into account during assessment as follows: "This regards the use of conjunctions, I mean, it is a grammatical concern. Why should I evaluate that separately?" The R3 replied to the question "Why don't you consider spelling as a component of assessment?" as follows: "Spelling mistakes do not interfere with the meaning. If existing spelling mistakes do not impede the comprehensibility of a text, I don't regard them as mistakes." Moreover, R3 reported that s/he paid attention to syntax too, yet s/he graded that within grammar component. As for task completion:

> "I don't think it is necessary to evaluate task completion in isolation. Students are naturally supposed to complete this and follow the instructions. If they do not obey the instructions, then they will be off topic, which has a certain fixed grade and there is no need to assess the essay. The highest I'll give to such a paper is 5 out of 100."

Assessing a total of seven criteria, R4 ignores only syntax, since s/he takes it as a part of grammar. The rationale for R5 not to assess task completion and syntax is that too many criteria lead to confusion. Therefore, s/he considers syntax as part of grammar and task completion as part of text features. So, the fact that raters take

various numbers of different criteria into account during assessment and neglect different benchmarks accounts for the weakness of intra-rater reliability emanating from assessment based on TDC.

Table 6 displays the steps of assessment systems described by each rater during interviews.

Table 6

Grading steps used by each rater

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
1	Reading each essay and making corrections on each paper (such as grammar mistakes), providing positive and negative feedback (such as "Good", "A nice remark", "Are you sure?")	Reading each essay by correcting grammar and spelling mistakes, grading either +10 or -10 in terms of task completion and text length.	Reading each essay and correcting grammar mistakes.	Reading each essay and grading them in terms of task completion (5 pts), text features (15 pts), and expression of ideas (15 pts).	Reading each essay and grading them in terms of grammar (25 pts) and vocabulary (25 pts).
2	Reading each essay and grading them by assigning a maximum of 50 points in terms of text features.	Re-reading each essay and grading them in terms of coherence/ cohesion (5 pts), and arguments and convincing effect (15 pts).	Grading arguments out of 25 pts, coherence /cohesion out of 25 pts, and vocabulary out of 25 pts.	Checking the corrections on the essays and grading them out of 15 for spelling and out of another 15 for grammar.	Re-reading each essay and grading them out of 25 for text features and out of 25 for coherence / cohesion.
3	Checking grammar mistakes and subtracting 1 point for each mistake (out of 15 pts) (repetitive grammar mistakes are subtracted only once).	Grading each paper out of 70 by subtracting ½ a point for each grammar, spelling, and wrong word use.	Re-reading each essay and checking grammar component out of 25 by grading 0 for each mistake and +1 point for each correct use.	Grading vocabulary out of 10 pts, and coherence /cohesion out of 20 pts.	Adding each point and grading each paper out of 100.
4	Grading Vocabulary out of 15.	Adding each point and grading each paper out of 100.	Adding each point and grading each paper out of 100.	Adding each point and grading each paper out of 100.	
5	Grading Coherence / Cohesion out of 15 pts.				
6	Adding each point and grading each paper out of 100.				

Analysis of scoring systems used by the participating raters reveals each rater has his/her own way of grading and that not only the benchmarks graded by the raters are different, but scoring of common benchmarks is also different across raters.

As for the steps involved in grading, R1 made use of six steps, while R2, R3, and R4 followed four steps, and R5 used a three-step grading system. Although the comparison of steps may show that the most time- and energy-consuming scoring system was employed by R5, the four-step scoring system, where R2 subtracted ½ a point for each grammar, spelling, and vocabulary mistake and then added up the points to reach a maximum of 70, should be noted as more time consuming and more tiring. Mistakes were scored with +/-1 or 0 by R1 and R3, too. However, R4 and R5 reported that they did not employ this technique, and they graded the papers in accordance with the set criteria and students' levels and through comparing students' performances.

R4 expressed his relevant ideas as follows:

> "I define my criteria and how to score them according to my students' levels; yet, sometimes I compare the essays with each other and then grade for a given benchmark. For example, if a B1 level student produces better grammar than his/her peers, then I give him/her more points than others. I believe s/he deserves that; I think that others should have also performed similarly."

With respect to common criteria employed during assessments, scores vary across different raters, and the significance of each criterion also differs across raters. For instance, "text features", a common criterion for all raters, was graded with the maximum grade 50 by R1, 15 by R2 and R4, and 25 by R3 and R5. So, R1 takes arguments and convincing discussion skills more seriously than others. Another common criterion, grammar, is graded with the maximum grade 15 by R1 and R4, and 25 by R3 and R5. How R2 graded this benchmark is not clear, because s/he graded grammar, spelling, and vocabulary as a whole out of 70. Lastly, vocabulary, as another common benchmark for all raters, is graded with the maximum grade 15 by R1, 25 by R3 and R5, and 10 by R4. Linguistic features do not matter for R3 and R5.

What Do the Raters Think about Using Rubrics?

Findings were classified in three categories, based on the analysis of semi-structured individual interviews conducted with five French language teachers/raters to learn their ideas about rubric use right after assessing 10 argumentative essays by scoring rubrics. These three categories were: The contributions of rubrics to the assessment, The contributions of rubrics to students and The contributions of rubrics to teachers.

Each category created as a result of analyzing the interviews was examined in sub-categories. Frequency values presented in Tables (Tables 7, 8, 9) displaying the findings indicate how often a situation/opinion was mentioned by the raters. Because one participant expressed his/her ideas repeatedly and about different topics, the total of frequencies exceeded the number of interviewees.

Table 7

Sub-categories regarding the contributions of rubrics to assessment procedure

Rubrics' contributions to assessment procedure	F
Ensure objectivity	16
Usability/practicability	15
Ensure validity	12
Total	43

As shown in Table 7, raters most often mentioned the contribution that use of rubrics made to the assessment procedure, namely objectivity. All 5 raters agree that criteria endowed within rubrics eliminated personal opinions and judgments. The following is what each rater said about this contribution:

> "I'm normally impressed by well-developed essays. When I read a good essay, I sometimes go back to other essays I had already graded and grade it again and sometimes tend to subtract some points because I think that all students are at the same level. So, if one of them can produce a complex sentence, the others must be able to do it, too.. If one of them makes use of a high-level expression, the others must do the same thing. But, rubrics prevented that. I believe I can assess each student independently from his/her peers. In short, I was quite objective with my assessment." (R4)

> "Today, I noticed something: earlier I used to assess papers based on only my criteria, my values, and within my understanding. Now I have a more limited space for my personal judgments." (R2)

> "The reason why I graded some essays with a lower mark earlier is that I valued grammar more than any other component. If there are many grammatical mistakes in an essay, I grow negative attitudes against that work … I mean I develop some kind of prejudice. For example, the student tries to explain one of his/her arguments and makes a serious grammatical mistake. I only focus on that mistake, and the content of the essay becomes invisible for me if there are too many grammatical mistakes. Yet, the use of rubrics stopped me this time. Grammatical mistakes did not pull the curtain over the good points in an essay." (R3)

> "Using rubrics is highly objective. No matter how hard I try to complete the assessment professionally, I always feel that there is some sort of subjectivity involved at all times, and that there is something missing…but it was different with rubrics. Of course, I can't say that using rubrics eliminates all sorts of problems, especially when one doesn't want to grade papers. Still, it offers a useful tool to conduct more objective assessment procedures." (R1).

Another frequent contribution of rubrics to assessment procedure is that they ease evaluating writing papers and provide a set of practical criteria. All raters

agree on this, and they noted the same thing about 15 times during the interviews. Participating raters stated that rubrics reduced the amount of time necessary to assess students' papers dramatically, and four of them (R1, R2, R3, R5) mentioned ease of interpretation and application, although there are many criteria within rubrics. The following are relevant quotations:

Time Reduction:

> "With rubrics, I had the opportunity to grade quickly and comfortably. Maybe the first paper took the longest since I used the rubrics for the first time, but then, I got used to them, and graded the rest of the works more quickly." (R1)
>
> "It is easier and faster to move along. A rater should carefully read and understand the rubrics before starting to assess." (R2)
>
> "At first, using rubrics is a little tiring because first you should read the rubrics carefully and understand and make sense of them. But, later it gets faster. I spent less time grading those papers. We both save time and manage a satisfying assessment procedure." (R4)
>
> "Time-saving. It gets faster after the first two papers. I can say that I'm faster now." (R5)
>
> "It doesn't matter how many papers there are. The system is all set. You keep grading accordingly. Grading according to rubrics is much faster even when there are many more papers." (R5)
>
> "Plus, we shouldn't forget how practical it is to grade a paper out of 25 and then multiplying the final grade by 4 in order to reach a score out of 100. It eases the job and speeds it up. This also affects the reduction of assessment time." (R2)
>
> "We are supposed to grade papers out of 100. Assessing them out of 25 and then multiplying the end score by 4 made things easier for us. I didn't feel tired, and I feel that the last score reflected how well an essay had been written. Seriously, much faster and easier." (R3)

Ease of application:

> "It was easier for me. The criteria within the rubrics are straightforward, clear, and to the point. It was so easy to assign the mark I had in my mind in accordance with the set of criteria. I mean, think about it; I graded various sub-skills, I evaluated them separately. I mean I graded writing papers and it wasn't difficult. All I had to do was to choose the score assigned to each benchmark." (R1)
>
> "You know grading writing papers is a tough and tiring job; this one wasn't like that; it was super easy to grade the papers. The reason for this was that everything is crystal clear. I read an essay, and I reflected on the content, grammaticality, and deficiencies of it through the use of rubrics. It is super easy to employ. All you have to do is to choose the corresponding score for each

benchmark; to do so, you have to check the rubric to see how good a student is at that piece of criterion." (R5)

Lastly, participating raters also reported validity as another contribution that the use of rubrics introduced to the assessment procedure. All raters agree that rubrics make it possible to evaluate writing performance appropriately and firmly, and that judgments about performances are realistic, or inferences made based on the performances are adequate. Quotations regarding this sub-category are as follows:

"It wasn't merely about grading a paper, it was more about assessing all relevant criteria that an argumentative essay should meet. I'm comfortable that the grade I gave matches the content quality of that essay… I don't have questions like: "Is the grade too high, or too low? Was that a good assessment?" … Normally, I have those questions…I often feel like I forgot or skipped something. Like, was I fair for all the papers…that's what I mean and…thanks to those graded criteria within rubrics." (R1)

"I was able to grade different aspects of essays. Total score I gave was good enough in reflecting the quality of an essay. I feel I assessed everything I was supposed to, and I did it appropriately and correctly. Assessment was to the point and as it was supposed to be… At least, I feel so." (R2)

"I had never employed such a systematic and organized method while evaluating writing papers. What to grade, and the maximum score I can give are all clear with rubrics. […] I was able to judge each criterion without making a comparison with others, and I believe I was really good at assessing papers, all thanks to rubrics. There are benchmarks to support the point I give, you know what I mean? er… the resulting grade, score, point, whatever you call it, is appropriate; there are things, criteria that validate and justify the grade I give." (R5)

Table 8

Sub-categories regarding the contributions of rubrics to students

Rubrics' contributions to students	f
Recognize areas that need improvement	8
Recognize what the writing expectations are	7
Total	15

According to the analysis of semi-structured interviews, raters noted that rubrics could also be beneficial for students, especially in terms of two aspects (Table 8): students can recognize the areas that need improvement and recognize what the writing expectations are. All raters referred to these two sub-categories a total of 15 times during the interviews. Instead of shortly mentioning these two gain areas, the raters provided detailed descriptions as to how rubrics can be beneficial for students in terms of these two sub-categories. As for the raters, students can conduct some sort of self-assessment about their writing performance and can learn what kind of

sub-skills are required and what they should improve to produce a qualified written product if they are provided with the rubrics. The following are quotations regarding "Recognize areas that need improvement":

> "I guess students can see how many and what kind of mistakes they made. They can judge what criterion they are good and bad at, and they can improve the weak points on the next writing task." (R3)
>
> "I think these rubrics contain more than points and scores. It looks like we use them to reach a final grade, but they also underline what students know and do not know. […] What if I hand these rubrics to students as I return their papers, and say: "Look, this is your paper, your score is 92. And these are the criteria I used while grading your paper. Have a look at them." Then, the student can clearly understand his/her mistakes and why s/he lost those 8 points. I won't have to make a detailed explanation. For example, the student can say "My vocabulary is poor." Accordingly, s/he will be more careful about vocabulary for the next task; s/he will try to improve his/her weak points." (R4)
>
> "When you give the rubrics to students, they will be able to analyze their mistakes. Having a closer look at the rubric will be enough for this. They can also compare the rubric and their papers, and conduct some kind of error analysis. During the next writing task, they will be more careful about their weak points they determined by studying the rubrics. At least, that's what I would do if I were a student." (R5)

Quotations are as follows regarding "Recognize what the writing expectations are":

> "Students will be able to see what is expected from them and what the writing skill requires if we give them the rubrics together with their papers. Yet, as far as I'm concerned… students should be given these rubrics at the beginning of the term and they should analyze them. Then, they can exactly understand what they are supposed to do … This will definitely help them… I would hand in the rubrics at the very beginning if it was up to me. Moreover, I would tell them to keep the rubrics, and use them while doing their homework assignments." (R1)
>
> "Indeed, they can understand the expectations, all requirements of writing task, and components of writing." (R3)
>
> "Students will be able to see what they need to do in order to write a good essay and what the components of a good essay are. They will be more aware of the things they should do to be successful. Their minds will be clear and sharp when they see the list of criteria in front of them. *[R4 asks a question to the researcher: I wanna ask something. Is it possible to adapt these rubrics for all levels? For A2 level for example? Researcher: Yes.]* Right, as students become more proficient, they can see that expectations change as well. I mean, if we grade some criteria higher than others for B2 level students, then students will notice and think that those criteria are more important. They can figure "The score for this benchmark is higher now, so I'm supposed to perform better on that." (R4)

Table 9
Sub-categories regarding the contributions of rubrics to teachers

Rubrics' contributions to teachers	f
Detailed/effective feedback	7
Use while teaching	7
Total	14

Raters also stated that rubrics can be beneficial for them as well. They noted that they could provide more-detailed feedback via use of rubrics, and they can employ them while teaching writing skills, too (Table 9). Of these two sub-categories, four raters (R1, R2, R4, R5) agreed with the first one, whereas three of them (R1, R2, R3) underlined the second one. The following are quotations regarding the sub-categories mentioned seven times by raters:

Detailed/Effective feedback:

> "Let me be honest, I have difficulty responding when students ask about their mistakes in their essays because I forget the contents of essays. Thus, I think these rubrics will be useful for me, too. When I check back the rubrics, I can easily say that student X is poor or successful in this and that. I believe rubrics are more explanatory. Plus, I can't provide feedback when there are many papers. I write more feedback at the beginning […] Students will not have to chase us to learn about their mistakes. Now, the amount of feedback on papers does not matter that much because rubrics are enough, they are also a kind of feedback, even a better kind." (R2)

> "I can easily explain why I scored X for any criterion via use of rubrics and I can do this for all the papers since they are assessed by the same criteria and under the same conditions. Er... Let me put it this way: sometimes, I hand out the papers I already graded to my students, and they ask for feedback regarding why they got a low grade, then I have to look back at the specific essay to remember the paper, and later I can say something about his work… The reason why I do so is that I sometimes do not punish students equally for the same mistakes, I mean, I give lower grades to some of them and don't do the same for others…er…that… When I give their papers back to them to let them see their mistakes, I tell them to analyze only their own papers. I tell them to attend to their own papers and to ask me if they have any questions without talking with their friends first. Yet, rubrics will put an end to this… I can comfortably explain exactly what is wrong with their performance and things that I took into consideration while assessing. I can provide consistent and reasonable explanations for each of them; I do not need to look back at their papers, it will be enough to look at the rubrics." (R4)

"Normally, I make 2 or 3 corrections on essays, and my students ask me if those few corrections were the reason for their low grades. It is not easy to give constant feedback. Yet, as soon as I give them the rubrics, each student will have his/her own feedback." (R5)

Use while teaching:

"If we study the rubrics all together… with the students I mean, in the classroom and through interaction… If we discuss what criterion X means and what criterion Y measures as a whole class activity, we will be teaching about writing as well. If we conduct some short activities regarding each criterion, […] really productive "coherence/cohesion" practice can be organized. I mean all items of rubric can be analyzed one by one: we can teach what it means to write successfully by using the rubrics as instructional material… Like, what is writing? This is what writing is… it is a skill consisting of the skills in the rubrics. We can state that argumentative essay involves the criteria defined by the rubrics, and study the rubrics." (R1)

"The criteria within the rubrics are actually the sub-skills necessary for writing skill … If we analyze each criterion one by one and explain them … talk about their function and importance… we'd be doing a writing class, I mean we'd be talking about the theory of writing… Theoretically, we can talk about what is necessary for a successful piece of writing task, the components of writing, and we can practice. At the end of the day, writing is a performance skill, true, but it is based on a relevant theoretical background… Normally, this theoretical part of writing class is ignored, and students start practicing directly. However, these rubrics offer an opportunity to teach about the theory of writing as well." (R2)

Discussion

There are numerous variables impeding effective assessment of performance (Black, 1998), and rater-related variable is the most frequently underlined one when it comes to assessing writing performance (Moskal, 2000). The reliability and validity of raters' judgments have often been questioned. In this regard, literature hosts an ample amount of studies suggesting use of rubrics, which is also the focal point of this research, to administer a positive assessment procedure and to produce more effective consequences.

The aim of the current study was to determine the reliability (intra/inter-rater) of rubrics after an assessment procedure, where five French language teachers who had never employed rubrics and always used their own criteria assessed 10 argumentative essays, first based on TDC and then on rubrics with a three-month interval between the ratings. Semi-structured interviews revealed what criteria varied across raters, and the quality of rubrics was examined.

In compliance with the first research question, the grades that participating raters assigned to essays, based on both TDC and on rubrics, were compared, and a dramatic

difference was identified between the grades. It was determined that some raters did not score the same paper with the same grade as a result of two different assessment techniques and that the disparity was as much as +/-16 for some papers. This meant that intra-rater reliability, the level of agreement between the two grades given by the same rater, was pretty low.

The second research question investigated inter-rater reliability by comparing the two grades given by two different raters to the same essay via use of different assessment procedures, and inter-rater consensus, the agreement level between raters, was analyzed. According to the research findings, inter-rater reliability was high for assessments through rubrics. Indeed, the grades that raters gave by using rubrics are very similar, indicating a sort of consensus. As for the assessments through TDC, a disagreement was identified between raters: they gave significantly different grades to the same papers when employing TDC.

The results obtained from the findings related to the first and second research questions were confirmed by statistical analysis, such as Spearman's rank correlation coefficient and Kendall's tau (W), and it was found that when raters use rubrics, they are more consistent and reliable than when using their self-made criteria. Such findings make sense, because rubrics serve as a grading guide for the raters by presenting them with all the criteria at a time. They also describe different quality levels of each criterion, which underscores how raters should assess the gist and content of a writing performance and what criteria to be careful about (Arter & McTighe, 2001; McMillan, 2004). Literature review yields a lot of studies indicating that scoring with rubrics is more consistent and reliable, that raters can reduce the variations between raters (inter-reliability), and that they can reduce the inconsistencies in the scoring process, due to raters' internal factors (intra-rater reliability) (Jonsson & Svingby, 2007; Hansson, Svensson, Strandberg, Troein, & Beckman, 2014; Moskal & Leydens, 2000).

In accordance with the qualitative data set collected during the semi-structured interviews conducted within the scope of the third research question, the criteria that raters employ when they apply TDC vary significantly, and the number of criteria used for assessment change dramatically: some raters make use of only four criteria, whereas others use six and seven criteria. Although three of the criteria (text features, grammar, and spelling) are common among the raters, grading benchmarks for these criteria also differ from rater to rater, which means that they do not agree on the value of a benchmark, even though they agree on some criteria. Besides, the steps they employ to read the essays are not the same either: whereas R1 spends a greater amount of time to grade one essay because s/he goes through six steps while assessing, R5 finishes the assessment of an essay in a shorter time, and s/he applies only two steps to finish the assessment of an essay. All these are explanatory enough as to what influences inter-rater reliability and why it is so low when TDC is employed during assessment. The variety of criteria on which raters establish their judgments and different benchmarks they assign to those criteria naturally impact the overall

score they give, leading to different scores for the same essay. As McNamara (1996, p. 117) stated, "performance assessment necessarily involves subjective judgments" and this subjectivity often influences differences in the type of rating criteria and scoring procedures or interpretation of rating criteria. Barkaoui (2007, p. 86) showed in his study that "raters were the main source of variability in terms of scores and decision-making behavior". Likewise, Schoonen (2005, p. 1) demonstrated that "the generalizability of writing scores and the effects of raters and topics are very much dependent on the way the essays are scored and the trait that is scored".

Finally, the fourth research question was directed to learn what participating teachers/raters who used rubrics for the first time while assessing writing papers thought about them, and a second set of interviews were held. Results have shown that raters especially underlined that use of rubrics produced more-objective scores and that using TDC led to subjective outcomes, focusing on fewer criteria and high appreciation of grammar. According to them, rubrics freed them from such mistakes causing subjectivity during assessment, because this assessment tool allowed them to focus only on the criteria by reflecting all criteria and their benchmarks as a whole: "Once I had the benchmarks of all criteria in front of me, I didn't have any difficulty focusing. I concentrated on each criterion, and complied with each of them" (R2). Ahoniemi and Reinikainen (2006, p. 139) support this opinion as well: "…the only way to […] still achieve objectivity is to divide the assessment in small enough parts with rubrics."

For a rubric, "the trait of practicality refers to ease of use" (Arter & McTighe, 2001, p. 49). Participating raters approached this issue from two different angles. First, it was faster to assess the papers, since rubrics presented all the criteria together. However, they also underlined that one had to comprehend the rubrics clearly and review each criterion carefully before the assessment. Some of them stated that they were rather slow grading the first essay, since they were also trying to understand the rubrics at the same time, and they got faster and faster as they assessed more papers. In addition, raters mentioned that it was practical to grade the essays out of 100 by using the rubrics, which also accelerated the assessment process. Considering R2, who subtracts ½ a point for each grammar or spelling mistake, and wrong word use in order to reach a total of 70 points when assessing through TDC, it is quite understandable how rubrics eased the assessment process. Similarly, rubrics are so useful for raters such as R1, who subtracts 1 point for each mistake to form a grammar score out of 15, and R3, who grades 0 for each mistake and 1 for each correct use in order to reach a total of 25 points for the grammar section. The second reason why it was easier and more practical to use rubrics according to the raters is that rubrics eased the assessment procedure without clouding the contents of papers: they were satisfied assessing each sub-skill precisely and easily by simply ticking small boxes. In their study, Bainer and Porter (1992, p. 12) explained that teachers agreed that using rubrics "to evaluate writing papers was easy because the rubric provided specific points to follow, thus providing a 'base on which to start'".

Another remark the raters made during the interviews regards the validity of the assessment procedure. According to Nitko (2004, p. 34), "To validate [their] interpretations and uses of students' assessments results, [raters] must provide evidence that these interpretations and uses are appropriate". This is exactly what participating raters stated: they noted that the clarity of criteria and the existence of graded scores for each criterion, made it easy for them to explain why they assign a specific grade for an essay comfortably. For instance, R4 made the following remark: "I looked at the criteria, then I assessed the paper. It is like my judgment about the quality of an essay had solid grounds… Total score reflected how good the work was and I felt that my judgment was appropriate and valid". Likewise, Bresciani et al. (2004, p. 30) stated that rubrics "combat accusations that evaluators do not know what they are looking for".

Lastly, raters noted that rubrics would also be beneficial for students as well as themselves. According to them, rubrics can permit learners to recognize areas that need improvement and what the writing expectations are, and they can permit teachers to make detailed feedback and teach about writing. As a matter of fact, the contributions reported by the participating raters for both themselves and the students are consistent with the results of many research studies. For instance, referring to a number of studies, Reddy and Andrade (2010, p. 437) noted: "When used by students as part of a formative assessment of their works in progress, rubrics can teach as well as evaluate". As for Stevens and Levi (2013), handing the rubrics to students is a very effective strategy to let them know what is expected of them, and it is really relaxing for many students to know the criteria beforehand. As far as Oakleaf is concerned (2009, p. 969), "rubrics allow students to understand the expectations of their instructors". Similarly, Bresciani et al. (2004, p. 30) stated that rubrics "make public key criteria that students can use in developing, revising, and judging their own work". Jaidev (2011, p. 7) also emphasized that rubrics have a crucial role in improving writing skills, and students can better express their opinions thanks to rubrics: "Knowledge of writing rubrics also helps students become more accountable for their own writing, and it allows them to gain a greater sense of ownership of what they have written". But because "rubrics are not entirely self-explanatory" (Andrade, 2005, p. 29), it may not be enough to plainly hand them out to the students and tell them to use the rubrics. "Students need help in understanding rubrics and their use" (Andrade, 2005, p. 29), and that is why teachers need to explain those to learners, to explain each criterion. Participating raters underpinned that studying the rubrics together with students would be a way of learning about writing and a tool for writing instruction. According to the participants, discussing and explaining the criteria in rubrics will help students either learn or retain the sub-skills of writing. As said by Arter and McTighe (2001, p. 10): "Clearly defined criteria and scoring guides provide more than just evaluation tools to use at the end of instruction – they help clarify instructional goals and serve as teaching targets". As for raters, the opportunity to provide more detailed feedback is another contribution of rubrics to assessment procedure. They noted that they were

only making few corrections and scratches on papers, since constant feedback is both tiring and time consuming, and that they were having difficulties explaining what their students' mistakes were and why they got that specific grade. They gladly stated that rubrics were of great use in terms of providing detailed, meaningful, and effective feedback by just looking at the rubrics and in terms of helping their students figure out what to do to improve their writing skills. Studies by Stevens and Levi (2005), Reddy and Andrade (2010), and Brookhart (2013) also support these findings.

Conclusion

In summary, it is possible to conclude that the use of rubrics during writing assessment produces more reliable and consistent outcomes, as indicated by both quantitative and qualitative findings of the current research. Rubrics are more credible and trustworthy, as they help the rater keep his/her judgments stable from one essay to another. Therefore, assessing writing papers through the use of tools containing certain criteria will probably eliminate inconsistencies among raters. Since raters' degree of leniency or severity is set at the beginning of the process, rubrics will require objectivity, as the criteria in them reflect instructional goals and only the gains to be measured, prevent addition of new criteria, and systematically focus on the same components (Berthiaume & Collet, 2013). Of course, it is not possible to reach 100% objectivity, due to the fact that the raters are human; so, there will always be the probability of subjectivity. Nevertheless, rubrics stand as one of the valuable tools to minimize subjectivity during performance assessment. Another feature that makes use of rubrics favorable is that students are not regarded as passive assessees; rather, they are actively integrated into the assessment procedure. The fact that participating raters noted that students need to study the rubrics together with their teachers can be taken as an indicator of this feature. Finally, it was determined that participating raters were not familiar with rubrics. Thus, it seems vital to integrate practical education/training about the use of rubrics into teacher training programs, so that teachers will not merely score a paper, but they will also be able to provide effective feedback and figure out the problems with writing skills.

References

Ahoniemi, T., & Reinikainen, T. (2006, February). ALOHA-a grading tool for semi-automatic assessment of mass programming courses. In *Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006* (pp. 139-140). https://doi.org/10.1145/1315803.1315830

Akbulut, Y. (2010). *Sosyal bilimlerde SPSS uygulamaları: Sık kullanılan istatiksel analizler ve açıklamalı SPSS çözümleri*. Istanbul: İdeal Kültür Yayıncılık.

Alpar, R. (2012). *Uygulamalı istatistik ve geçerlik-güvenirlik: spor, sağlık ve eğitim bilimlerinden örneklerle*. Ankara: Detay Yayıncılık.

Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College teaching*, *53*(1), 27-31. https://doi.org/10.3200/CTCH.53.1.27-31

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. California: Corwin Press.

Bainer, D., & Porter, F. (1992, October). *Teacher concerns with the implementation of holistic scoring*. Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86-107. https://doi.org/10.1016/j.asw.2007.07.001

Berthiaume, D., & Rege Colet, N. (2013). Comment développer une grille d'évaluation des apprentissages ? In D. Berthiaume, & N. Rege Colet (Eds.), *La pédagogie de l'enseignement supérieur : repères théoriques et applications pratiques* (Vol. 1, pp. 269-283). Berne: Peter Lang. https://doi.org/10.3726/978-3-0352-0230-4

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

Bresciani, M. J., Zelna, C. L., & Anderson, J. A. (2004). *Assessing student learning and development. A handbook for practitioners.* United States: NASPA.

Breton, G., Lepage, S., & Rousse, M. (2010). *Réussir le DELF B1-CIEP.* Paris: Didier.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Virginia: Ascd.

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, *67*(3), 343-368. https://doi.org/10.1080/00131911.2014.929565

Can, A. (2014). *SPSS ile bilimsel araştırma sürecinde nicel veri analizi* (2nd ed.). Ankara: Pegem Akademi.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, *14*(2), 88-115. https://doi.org/10.1016/j.asw.2009.04.001

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, *12*(1), 1-9. https://doi.org/10.1016/j.asw.2007.05.002

Hansson, E. E., Svensson, P. J., Strandberg, E. L., Troein, M., & Beckman, A. (2014). Inter-rater Reliability and Agreement of Rubrics for Assessment of Scientific Writing. *Education*, *4*(1), 12-17.

Jaidev, R. (2011). Rubrics-based Writing: Liberating rather than Restricting in Many Contexts. *ELT World Online, 3*, 1-7. Retrieved from http://blog.nus.edu.sg/eltwo/files/2014/06/Rubrics-based-Writing_editforpdf-1a0neat.pdf

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

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

McMillan, J. H. (2004). *Classroom assessment: Principles and practice for effective instruction* (3rd ed.) Boston: Pearson.

McNamara, T. (1996). *Measuring second language performance*. New York: Addison Wesley.

Moskal, B. M. (2000). Scoring rubrics: what, when and how?. *Practical Assessment, Research & Evaluation,* 7(3). Retrieved from http://pareonline.net/getvn.asp?v=7&n=3

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*. Retrieved from http://pareonline.net/getvn.asp?v=7&n=10

Nitko, A. J. (2004). *Educational Assessment of Students* (4th ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

Oakleaf, M. (2009). Using rubrics to assess information literacy: An examination of methodology and interrater reliability. *Journal of the American Society for Information Science and Technology*, *60*(5), 969-983. https://doi.org/10.1002/asi.21030

Özçelik, D. A. (2010). *Ölçme ve değerlendirme*. Ankara: Pegem Akademi.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Romainville, M. (2011). Objectivité versus subjectivité dans l'évaluation des acquis des étudiants. *Revue internationale de pédagogie de l'enseignement supérieur*, *27*(2), 1-9.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. Bruxelles: De Boeck Supérieur.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1-30. https://doi.org/10.1191/0265532205lt295oa

Simard, C. (1992). L'écriture et ses difficultés d'apprentissage. In R. Ouellet, & L. Savard (Eds.), *Pour favoriser la réussite scolaire* (pp. 276-234). Montréal: Editions Saint-Martin.

Stevens, D. D., & Levi, A. J. (2005). *Introduction to Rubrics.* Sterling, VA: Stylus Press.

Tompkins, G. E. (2004). *Teaching writing: Balancing process and product*. (4th ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

Veda Aslim-Yetis
Anadolu University, Faculty of Education
Department of Foreign Language Education
26470 Eskisehir, Turkey
vaslim@anadolu.edu.tr

Appendix

Rubric for argumentative essay (English version) – B1 – 25 points

Following instructions									
Is able to apply his/her writing skills to the situation proposed. Is able to follow the instruction provided regarding minimum lenght.	0	0.5	1	1.5	2				
Ability to present facts									
Is able to describe facts, events, and experiences.	0	0.5	1	1.5	2	2.5	3	3.5	4
Ability to express thoughts									
Is able to present his/her ideas, feelings, and/or reactions and give his/her opinion.	0	0.5	1	1.5	2	2.5	3	3.5	4
Coherence and cohesion									
Is able to connect a series of short, simple, distinct elements in a discourse that flows.	0	0.5	1	1.5	2	2.5	3		

Lexical competence / Lexical spelling

Vocabulary range					
Has sufficient vocabulary to write about current topics, paraphrasing if necessary.	0	0.5	1	1.5	2
Vocabulary control					
Demonstrates good control of basic vocabulary, but major errors still occur when expressing more complex thoughts.	0	0.5	1	1.5	2
Orthographic control					
Lexical spelling, punctuation, and layout are accurate enough to be followed easily most of the time.	0	0.5	1	1.5	2

Grammatical competence / Grammatical spelling

Degree of elaboration in sentence structure					
Good control of simple sentence structures and more common complex sentence structures.	0	0.5	1	1.5	2
Choice of tense and mood					
Demonstrates good control though with noticeable mother tongue influence.	0	0.5	1	1.5	2
Morphosyntax – Grammatical spelling					
Agreement in gender and number, pronouns, verb endings, etc.	0	0.5	1	1.5	2

Procjena vrednovanja eseja: kriteriji vrednovanja koje su izradili nastavnici i rubrike. Individualna i međusobna pouzdanost ocjenjivača i mišljenja nastavnika o tome

Sažetak

Pouzdanost ocjenjivača ima ključnu ulogu u vrednovanju eseja, koje mora biti valjano, pouzdano i učinkovito. Ciljevi ovoga istraživanja su: odrediti individualnu i međusobnu pouzdanost ocjenjivača na temelju dviju skupina ocjena koje je pet nastavnika/ocjenjivača dalo tijekom vrednovanja 10 raspravljačkih eseja koje su napisali studenti koji uče francuski jezik kao strani jezik, a vrednovanje se provodilo u skladu s kriterijima koje su nastavnici sami izradili i uz pomoć rubrika; razumjeti kriterije kojima su se koristili u procesu vrednovanja; zabilježiti što ocjenjivači/nastavnici koji su se prvi put u procesu vrednovanja koristili rubrikama misle o takvom načinu vrednovanja. Kvantitativni podaci pokazali su da je individualna pouzdanost ocjenjivača s obzirom na ocjene koje su dali na temelju vlastitih kriterija vrednovanja i na temelju rubrika niska; da je međusobna pouzdanost ocjenjivača niska i kada se radi o ocjenama na temelju vlastitih kriterija te da je međusobna pouzdanost ocjenjivača veća u procesu vrednovanja uz pomoć rubrika. Kvalitativni podaci dobiveni putem metode individualnih intervjua pokazali su da su se ocjenjivači koristili različitim kriterijima. Tijekom drugoga kruga individualnih intervjua nakon primjene rubrika ocjenjivači su primijetili da su im rubrike pomogle u postizanju veće objektivnosti, da su pozitivno utjecale na proces vrednovanja i da se mogu koristiti kako bi pomogli studentima u procesu učenja i kako bi poboljšali provedbu nastavnoga procesa.

Ključne riječi: *mješovit model istraživanja; pisanje; procjena.*

Uvod

Nastava stranih jezika usmjerena je prema poboljšanju četiriju osnovnih vještina (čitanja, pisanja, govorenja i slušanja) i njihovih važnih podvještina. Pisanje, kao

jedna od osnovnih vještina, zahtijeva primjenu brojnih podvještina kako bi se poruka uspješno prenijela. Ta vještina ne zahtijeva samo točnu uporabu jezičnoga znanja (gramatike, vokabulara, točnoga pisanja riječi, itd.) nego i stvaranje originalnih ideja, organiziranje tih ideja u razumljiv oblik, jasno izražavanje misli, poticanje interesa kod čitača i razumljivost. Drugim riječima, učenici moraju imati kontrolu nad svojim vlastitim tekstom u smislu sadržaja, stila, organizacije ideja, vrste teksta (opisivanje, rasprava), jezičnih zakonitosti ciljnoga jezika i konvencija (interpunkcija, upotreba velikoga i malog slova, podjela u odlomke itd.). Isto tako, nastavnici moraju pažljivo analizirati sve podvještine koje pisanje podrazumijeva i razumjeti brojne informacije dok istovremeno „prosuđuju o kvaliteti – u kojoj je mjeri ponašanje ili rad učenika dobar" (McMillan, 2004, str. 10). Stoga je vrednovanje pisanih radova učenika jako naporno i zamorno te nastavnicima oduzima puno vremena. Zbog toga je „..već duži niz godina, […] vrednovanje pisanih uradaka preplavljeno pitanjima o pouzdanosti ocjenjivanja (što se obično odnosi na pouzdanost ocjenjivača)" (Hamp-Lyons, 2007, str. 1).

> Pouzdanost se odnosi na dosljednost ocjena danih u procesu vrednovanja. Na primjer, na pouzdanome testu učenik može očekivati istu ocjenu bez obzira na to kada je završio test, kada je test ocijenjen te tko ga je ocijenio (Moskal i Leydens, 2000).

Objektivno vrednovanje pisanih uradaka, ili barem što je moguće objektivnije ocjenjivanje, iznimno je važno. Mogućnost da procjena točnosti dugih odgovora može varirati od ocjenjivača do ocjenjivača dovodi u pitanje pouzdanost i valjanost vrednovanja (Ozcelik, 1992, str. 127). Vrednovanje eseja na temelju objektivnih kriterija svodi mogućnost pogrešaka u procesu vrednovanja na minimum i vodi k nepristranim i točnim rezultatima o učeničkim vještinama. Međutim, nije moguće potpuno eliminirati pogreške iz procesa vrednovanja. Kako je učenje apstraktan proces i kako nije moguće izravno vrednovati vještinu pisanja, primjenjuju se oblici neizravnog vrednovanja kako bi se prikupili podaci o vještini pisanja kod učenika. K tomu, na proces vrednovanja negativno mogu utjecati i pogreške koje se javljaju kod učenika, ocjenjivača, alata i metoda koje se koriste u vrednovanju. Na primjer, nekoliko faktora, poput ocjenjivanja pisanih uradaka u različito vrijeme, fizičkoga (umor) i psihičkog (radost ili tuga) stanja te vremena kada se vrednovanje provodi (prerano ili prekasno) može rezultirati pogreškama u procesu vrednovanja. Čak i pozitivne i negativne emocije ocjenjivača prema učenicima mogu utjecati na njega i dovesti do pogrešaka u vrednovanju. Stoga vrednovanje pisanoga uratka od istoga ocjenjivača u različitom vremenu može rezultirati različitim ocjenama, a učenici će morati živjeti s drastično različitim ocjenama, iako je njihov pisani uradak jednak. Nedosljednosti koje se javljaju zbog unutarnjih i vanjskih čimbenika tijekom procesa vrednovanja odnose se na individualnu pouzdanost ocjenjivača (Moskal i Leydens, 2000).

Nadalje, različite ocjene mogu biti i rezultat upotrebe različitih kriterija ili različitoga vrednovanja istih kriterija tijekom vrednovanja pisanih uradaka. Na primjer, nekim

ocjenjivačima važnije su gramatika i gramatičke strukture, a drugim su ocjenjivačima bitni uporaba jezika i debatne vještine (Moskal, 2000). Ipak, neki ocjenjivači procjenjuju rad učenika na temelju teksta kojega je napisao, dok drugi procjenjuju isti rad usporedbom s radovima drugih učenika (Romainville, 2011). Zato se pouzdanost procesa vrednovanja smanjuje i udaljava od objektivnosti. Te nedosljednosti, koje se javljaju zbog specifičnih i različitih kriterija vrednovanja ocjenjivača, pripadaju području međusobne pouzdanosti ocjenjivača (Moskal i Leydens, 2000).

Ono što se od ocjenjivača očekuje jest provedba pouzdanoga vrednovanja koje će rezultirati sličnim ili istim ocjenama, a na koje neće utjecati osobna prosudba, čimbenici izvan područja nastavnih ciljeva, vremenska i prostorna ograničenja. Kako bi se to postiglo, potrebno je unaprijed odrediti kriterije vrednovanja i vrednovati sve učeničke uratke u skladu s tim kriterijima.

Analiza postojeće relevantne literature pokazala je da su rubrike za ocjenjivanje jedan od učinkovitih alata kojima se ocjenjivači mogu koristiti kako ne bi zanemarili kriterije vrednovanja, kako ne bi (svjesno ili nesvjesno) mijenjali kriterije od jednog eseja do drugog te kako bi pokazali kriterije vrednovanja kao cjelinu za vrednovanje uspjeha učenika u određenom području, kao što je vještina pisanja (Berthiaume i sur., 2011; Brookhart i Chen, 2015; East, 2009; Jonsson i Svingby, 2007; Moskal, 2000; Moskal i Leydens, 2000; Nitko, 2004; Scallon, 2004; Reddy i Andrade, 2010). Tako i Moskal (2000) navodi: „Rubrike za ocjenjivanje obično se koriste kada se zahtijeva procjena kvalitete, a mogu se koristiti i za evaluaciji raznovrsnih tema i aktivnosti." Andrade (2005, str. 27) ovako definira rubriku:

> „Rubrika je alat za vrednovanje koji navodi kriterije za određenu vrstu uratka ili navodi ono što je za takav uradak bitno (na primjer, svrha, organizacija, detalji, glas i tehnika su ono što je u pisanome eseju bitno) i prikazuje gradaciju kvalitete za svaki kriterij, od odlične do slabe."

Rubrike za ocjenjivanje izrađene su u skladu s određenim kriterijima i za svaki kriterij određuju što se ocjenjuje. One su stupnjevit i opisn alat za ocjenjivanje koji se koristi kako bi se zadržalo standardno i stabilno vrednovanje. Taj alat može biti ili u već gotovome obliku ili ga mogu izraditi nastavnici/ocjenjivači, na temelju kvalitete pisanoga uratka (Stevens i Levi, 2005). Rubrike mogu biti dvojake: holističke i analitičke. Holističke rubrike „ocjenjuju ili procjenjuju rezultat ili proces kao cjelinu, bez prethodnog zasebnog ocjenjivanja dijelova" (Nitko, 2004, str. 264). U rasponu od 3 boda do 6 bodova, holističko ocjenjivanje podrazumijeva brzo i lako vrednovanje, što je čest izbor nastavnika koji rade u razredima s velikim brojem učenika. Takva vrsta vrednovanja uglavnom je usmjerena na konačni proizvod/rezultat, a ne na postupak koji je možda imao utjecaj na učenikov rad. Na primjer, rubrika za holističko vrednovanje izrađena za vrednovanje pisanih uradaka učenika temelji se na skali od 1 do 3: Odličan (3 boda) – jasno izražavanje misli, mišljenje potkrijepljeno primjerima,

nema pogrešaka u pisanju riječi, nema gramatičkih pogrešaka; Dobar (2 boda) – misli su razumljivo prezentirane, primjeri ne odgovaraju u potpunosti iznesenom mišljenju, riječi su uglavnom točno napisane, ima malo gramatičkih pogrešaka; Slab (1) – misli nisu jasno izložene, mišljenje nije potkrijepljeno primjerima, vidljive su mnoge pogreške u pisanju riječi, postoji puno gramatičkih pogrešaka. U skladu s takvim načinom ocjenjivanja rad učenika može se ocijeniti ocjenom od 1 do 3. Međutim, navedene rubrike ne pružaju odgovarajuću povratnu informaciju koja bi učenicima pomogla u poboljšanju vještine pisanja i ne omogućavaju zasebnu analizu svakoga kriterija jer je pristup svim kriterijima holistički. Stoga su sve češća istraživanja koja pokazuju da je analitičko, a ne holističko vrednovanje eseja prikladnije i učinkovitije za vrednovanje vještine pisanja (Brookhart, 2013; Lumley, 2002).

„Analitičke rubrike opisuju rad na svakome kriteriju zasebno" (Brookhart, 2013, str. 6). „Te rubrike najprije ocjenjuju ili procjenjuju zasebne dijelove ili obilježja uratka ili procesa, a zatim se pojedinačne ocjene zbrajaju kako bi se došlo do ukupne, konačne ocjene" (Nitko, 2004, str. 264). Stoga, kada se koriste za vrednovanje vještine pisanja, te rubrike procjenjuju svaku komponentu pisanoga uratka zasebno i stupnjevito te omogućavaju detaljnu procjenu svakoga dijela. Na primjer, rubrike se fokusiraju na nekoliko tema, kao što su: izvršenje zadatka, sadržaj, izražavanje misli, izbor riječi i gramatika. Za svaku od tih tema postoji ključ za stupnjevito ocjenjivanje. Pisani uradak može dobiti ocjenu 4 za sadržaj, ocjenu 5 za izbor riječi i gramatiku, ocjenu 2 za sintaksu. Nastavnici dobro znaju kako ocijeniti svaki kriterij, a učenici mogu lako uvidjeti koji im je kriterij slabiji i znaju da na njemu moraju raditi i eliminirati takve pogreške na sljedećem zadatku pisanja. Drugim riječima, s pomoću takvih rubrika prepoznaju se jake i slabe podvještine učenika, dokumentira nastavnikov način vrednovanja i daje se detaljnija povratna informacija o slabije razvijenim vještinama učenika, umjesto da nastavnik samo napiše nekoliko nejasnih riječi na pisane uratke učenika. Kako su naveli Jonsson i Svingby (2007, str. 132): „Analitičko ocjenjivanje korisno je u razredu jer rezultati i nastavnicima i učenicima pomažu da prepoznaju jake strane kod učenika, kao i njihove potrebe."

Glavni je cilj ovoga istraživanja odrediti pouzdanost rubrika pri ocjenjivanju uspjeha učenika u vještini pisanja i prepoznati njihov utjecaj na individualnu i međusobnu pouzdanost ocjenjivača. U skladu s tim postavljena su sljedeća pitanja istraživanja:

1. Postoje li razlike u individualnoj pouzdanosti ocjenjivača u vrednovanju eseja s pomoću kriterija koje su izradili sami nastavnici i s pomoću rubrika za ocjenjivanje?
2. Postoje li razlike u međusobnoj pouzdanosti ocjenjivača u vrednovanju eseja s pomoću kriterija koje su izradili sami nastavnici i s pomoću rubrika za ocjenjivanje?
3. Kojim se kriterijima vrednovanja i načinom ocjenjivanja koriste nastavnici/ocjenjivači koji su sudjelovali u istraživanju?
4. Što ocjenjivači misle o služenju rubrikama?

Metodologija

Model istraživanja

U ovom se istraživanju koristio sekvencijalni eksplanatorni dizajn, kao vrsta istraživanja u kojemu se koriste mješovite metode za prikupljanje i analiziranje kvalitativnih i kvantitativnih podataka, kako bi se došlo do odgovora na postavljena pitanja istraživanja. Ovakva vrsta istraživanja započinje s prikupljanjem i analizom kvantitativnih podataka koji bi mogli dati odgovore na pitanja istraživanja, a zatim se nastavlja dubinskom analizom kvantitativnih podataka primjenom kvalitativnih podataka kako bi ih se moglo protumačiti (Creswell i Plano Clark, 2011).

U ovome istraživanju kvantitativni podaci prikupljeni su putem vrednovanja pisanih uradaka i u skladu s kriterijima koje su izradili nastavnici i uz pomoć rubrika, a kvalitativni podaci prikupljeni su putem polustrukturiranih intervjua.

Sudionici

U ovom istraživanju sudjelovalo je pet nastavnika koji predaju francuski jezik u višim školama za strane jezike povezanima s dva sveučilišta u Turskoj. Nastavnici su u istraživanju sudjelovali dobrovoljno.

Tri glavna kriterija odabira kandidata bila su: barem dvije godine radnoga iskustva u nastavi francuskoga jezika, rad na kolegiju Pisanje i nikakvo prethodno iskustvo u radu s rubrikama. Iz etičkih razloga nisu navedena imena nastavnika, kako bi se zaštitila njihova privatnost. Ovo su opće karakteristike nastavnika/ocjenjivača:

Ocjenjivač 1: star 30 godina; predaje francuski jezik u Višoj školi za strane jezike pet godina; drži kolegije „Čitanje“ i „Pisanje“; ima iskustva u podučavanju vještine govora; prethodno nije radio ni u jednoj drugoj instituciji.

Ocjenjivač 2: star 27 godina; predaje francuski jezik u Višoj školi za strane jezike dvije godine; drži samo kolegij „Pisanje“ i prije nije radio ni u jednoj instituciji.

Ocjenjivač 3: star 28 godina; predaje francuski jezik u Višoj školi za strane jezike četiri godine; drži kolegij „Jezične aktivnosti“ u kojemu se, uz pisanje, provode različite aktivnosti; prethodno nije radio ni u jednoj drugoj instituciji.

Ocjenjivač 4: star 32 godine; predaje francuski jezik u Višoj školi za strane jezike četiri godine; drži kolegije: „Čitanje“, „Govorenje“, „Slušanje“ i „Pisanje“; prethodno nije radio ni u jednoj drugoj instituciji.

Ocjenjivač 5: star 35 godina; predaje francuski jezik u Višoj školi za strane jezike tri godine; drži kolegije „Čitanje“ i „Pisanje“; prethodno nije radio ni u jednoj drugoj instituciji.

Instrumenti/alati za prikupljanje podataka

Eseji

Eseje koji su ocijenjeni u sklopu ovoga istraživanja napisalo je 10 studenata (na B1 razini) koji uče francuski jezik kao strani jezik. Svaki je student napisao jedan

raspravljački esej kao dio završnoga ispita u kolegiju „Pisanje". Razlog zašto je istraživač odabrao eseje pisane na ispitu jest pretpostavka da će studenti puno ozbiljnije pristupiti zadatku pisanja. Svi studenti obaviješteni su o istraživanju, a prije njegove provedbe dali su svoj pristanak.

Studenti su dobili ovakvu uputu za pisanje raspravljačkoga eseja: „*Mislite li da je dobra ideja obrazovati se u inozemstvu?* Izrazite svoje mišljenje i potkrijepite ga argumentima. Napišite jasan i koherentan esej čiji tijek, organizacija i stil odgovaraju zadatku, svrsi i čitateljima. Vrijeme: 90 minuta. Broj riječi: 160 – 180." Razlog odabira raspravljačkoga eseja jest taj što on od studenata zahtijeva dosljednu prezentaciju realističnih ideja utemeljenih na stvarnosti. Pri pisanju takvoga eseja glavna je kognitivna odgovornost studenta ne pisati o nestvarnim idejama, nego navesti stvarne, istinite informacije. Raspravljački esej treba čitatelje uvjeriti u nešto, nametnuti ideju i potaknuti ih da nešto učine ili ne učine. Tako autor mora tražiti i prezentirati razumne i uvjerljive argumente (Tompkins, 2004, str. 421).

20 studenata bilo je prisutno na ispitu, ali je istraživanje ograničeno na broj od 10 studenata kako bi se smanjio broj mogućih pogrešaka koje bi mogle utjecati na proces vrednovanja. U stvari, tijekom prijašnjih intervjua pet je ocjenjivača smatralo da je broj eseja razuman i da se mogu lako ispraviti. 10 eseja odabrano je na sljedeći način: analizirano je svih 20 eseja koje je napisalo 20 studenata koji su pohađali kolegij iz pisanja i koji su izašli na završni ispit. Odabrano je 10 eseja s obzirom na broj ideja, njihovu dosljednu prezentaciju i duljinu eseja.

Polustrukturirani individualni intervjui tijekom prve faze

Cilj polustrukturiranih intervjua koji su održani odmah nakon što su ocjenjivači ocijenili eseje na temelju vlastitih kriterija bio je utvrditi koje su kriterije vrednovanja ocjenjivači uzeli u obzir ili ih nisu smatrali bitnima, opisati pojedinačne teorije i konstrukte na temelju kojih su ocjenjivači temeljili proces vrednovanja te razumjeti njihov sustav ocjenjivanja.

U skladu s tim ocjenjivačima su postavljena pitanja otvorenoga tipa kako bi se došlo do spoznaja koji su im kriteriji bili bitni, a koje su zanemarili tijekom procesa vrednovanja. Štoviše, ocjenjivače se podsjetilo na druge kriterije koje nisu uzeli u obzir i pitani su zašto je to bilo tako.

Analitička rubrika

U ovome istraživanju koristila se analitička rubrika na francuskom jeziku koja je izrađena za vrednovanje raspravljačkih eseja. Ukupan broj bodova je 25, a rubriku je izradio Međunarodni centar za pedagoške znanosti kako bi se vrednovali raspravljački eseji učenika na B1 razini (Breton i sur., 2010, str. 74). Rubrika se sastoji od tri dijela: prvi dio obuhvaća stavke o izvršenju zadatka, koherenciji, koheziji i jasnom izlaganju ideja (ukupno 13 bodova); drugi dio sastoji se od dijelova koji se odnose na leksičku kompetenciju i pravilno pisanje riječi (ukupno 6 bodova), a uključuje

i raspon vokabulara i pravopis; treći dio usmjeren je na gramatičku kompetenciju (ukupno 6 bodova) i sadrži dijelove o strukturi rečenice, glagolskom vremenu/načinu i morfosintaksi.

Ukratko, prvi dio procjenjuje mogu li se dijelovi raspravljačkoga eseja navedeni u uputi prepoznati ili ne (npr. broj riječi, izlaganje argumenata itd.) i vrednuje načela vještine pisanja, a drugi i treći dio uglavnom procjenjuju poznavanje gramatike i vokabulara koje doprinosi boljoj kvaliteti pisanoga uratka. Kako je naveo Simard (1992, str. 286), pisac, za razliku od govornika, mora biti jasniji, razumljiviji i eksplicitniji kako bi izrazio svoje ideje, koristeći se „širim i raznovrsnijim lingvističkim repertoarom". U samoj rubrici navedena su i objašnjenja svake stavke (vidi Prilog), a kako bi došli do ocjene 100, ocjenjivači su morali zbrojiti ukupan rezultat bodova iz rubrika i jednostavno ga pomnožiti s brojem 4.

Polustrukturirani individualni intervjui

Polustrukturirani intervjui provedeni su odmah nakon što su ocjenjivači ocijenili svih 10 eseja (koje su vrednovali bez rubrika tri mjeseca prije) uz pomoć analitičkih rubrika. Svrha tih intervjua bila je utvrditi mišljenja ocjenjivača o primjeni rubrika, je li im taj način vrednovanja bio praktičan ili nije te koje negativne i pozitivne aspekte primjene rubrika mogu navesti.

Praksa/provedba

Prije istraživanja nastavnici francuskoga jezika u Višoj školi za strane jezike kontaktirani su i upoznati s istraživanjem, a zatim je odabrano pet nastavnika koji su odgovarali zadanim kriterijima odabira. Njima su dane detaljne informacije o istraživanju, kao što je dvostruko ocjenjivanje 10 eseja koje su napisali studenti koji uče francuski jezik. Isto su tako upoznati s činjenicom da njihovi osobni podaci neće biti objavljeni te da se njihova imena neće povezati s komentarima koje će navesti.

Nakon toga je svaki nastavnik ocijenio 10 eseja u različite dane, koristeći se vlastitim kriterijima. Eseji su numerirani nasumično, a nastavnici su dobili upute da ocijene eseje istim redoslijedom. Kako raspon ocjena na turskim sveučilištima doseže broj 100, nastavnicima je rečeno da eseje i ocijene ocjenom do 100. Ocjenjivači su bili sami u sobi, kako bi se smanjilo odvraćanje pažnje i kako bi se izbjeglo ometanje tijekom ocjenjivanja. Odmah nakon ocjenjivanja s ocjenjivačima su provedeni polustrukturirani individualni intervjui.

Istraživači su smatrali da je vremenski razmak od tri mjeseca dovoljan kako se ocjenjivači više ne bi sjećali sadržaja eseja, niti kako su ih ocijenili, te da se može provesti druga faza istraživanja. Tri mjeseca poslije ocjenjivači su ponovno pozvani u različite dane. Ovaj su put trebali ocijeniti iste raspravljačke eseje koristeći se nizom analitičkih rubrika. Osim toga, rečeno im je da slijede isti redoslijed prema brojevima na esejima, što je pomoglo da se zadrži isti redoslijed ocjenjivanja i bez pomoći analitičkih rubrika i uz njihovu pomoć. Ocjenjivanje se provodilo ocjenama do

100. Slično kao i u prvoj fazi ocjenjivači su vrednovanje provodili sami u zasebnoj prostoriji, ali je prije samoga vrednovanja sa svakim ocjenjivačem održan kratak sastanak (20 minuta) kako bi se porazgovaralo o rubrikama koje će upotrijebiti i kako bi se razjasnile nejasnoće o rubrikama. Odmah nakon vrednovanja sa svakim je ocjenjivačem proveden polustrukturirani intervju.

Rezultati

Postoje li razlike u individualnoj pouzdanosti ocjenjivača u vrednovanju eseja s pomoću kriterija koje su izradili sami nastavnici i s pomoću rubrika za ocjenjivanje?

Tablica 1 prikazuje razliku u ocjenama koje su ocjenjivači dali svakom eseju koristeći se kriterijima koje su sami izradili i koristeći se rubrikama, kao i broj eseja unutar istoga raspona ocjena i informacije o individualnoj pouzdanosti ocjenjivača.

Tablica 1

Detaljnija analiza Tablice 1 pokazuje da nijedan ocjenjivač nije jednako ocijenio eseje koje su već prethodno ocijenili. Ocjenjivač 1 (O1) ima najmanju podudarnost u ocjenama, jer je razlika u njegovim ocjenama bila $\pm$16 u pet od 10 eseja.

Slično tome, O5 je drugi po redu po najmanjoj podudarnosti u ocjenama, jer je razlika u njegovim ocjenama bila $\pm$16 u četiri od 10 eseja koje je ocijenio. Dok su O1, O2, O3 i O4 dali ocjene s razlikom $\pm$1 – 5, O5 nije nijedan esej ocijenio u takvom rasponu te su njegove ocjene imale razliku $\pm$6 – 10 u pet eseja. Na temelju rezultata prikazanih u tablici može se jednostavno zaključiti da su ocjenjivači dali ocjene koje su raznolike te da je njihova individualna pouzdanost niska. Koeficijent korelacije između vrednovanja s pomoću rubrika i bez njih analiziran je kako bi se bolje shvatila pouzdanost ocjenjivača u dvije vrste vrednovanja te kako bi se moglo doći do statističke definicije. Zbog toga se koristio Spearmanov koeficijent korelacije ranga – neparametrijski ekvivalent Pearsonova koeficijenta korelacije (produkt moment korelacije) (Tablica 2).

Tablica 2

Kako se može vidjeti u Tablici 2, koeficijenti korelacije variraju u rasponu od - 0,016 i 0,631, a nijedna od tih vrijednosti nije statistički značajna. To znači da su ocjenjivači različito ocijenili istih 10 eseja kada su se koristili rubrikama. Stoga se može reći da ocjenjivači nemaju dosljedan sustav ocjenjivanja i da je individualna pouzdanost ocjenjivača u vrednovanju pisanih uradaka prilično niska. Štoviše, Tablica 3 pokazuje da razlika u ocjenama koje su dane uz pomoć kriterija koje su nastavnici izradili sami i uz pomoć rubrika varira između 2 i 38. Na primjer, O3 dao je ocjenu 88 četvrtom eseju, koji je prethodno ocijenio ocjenom 50 (razlika od +38). Isto tako, O5 dao je ocjenu 92 osmome eseju, koji je prije ocijenio ocjenom 55 (razlika od +37). Isti je ocjenjivač sedmi esej ocijenio ocjenom 90, a prethodno ga je ocijenio ocjenom 68 (razlika od +22). O5 je s pomoću samostalno izrađenih kriterija vrednovanja ocijenio treći esej

ocjenom 90, a poslije ga je uz pomoć rubrika ocijenio ocjenom 66 (razlika -22). Kod svih ocjenjivača mogle su se uočiti manje i veće razlike u ocjenama.

Tablica 3

Postoje li razlike u međusobnoj pouzdanosti ocjenjivača u vrednovanju eseja s pomoću kriterija koje su izradili sami nastavnici i s pomoću rubrika za ocjenjivanje?

Tablica 3 pokazuje da ne postoje razlike samo u ocjenama kojima je isti ocjenjivač ocijenio isti esej, ovisno o alatu koji se koristio za vrednovanje, nego i u ocjenama koje su različiti ocjenjivači dali istome eseju. Na primjer, ocjene za treći, peti i sedmi esej variraju u rasponu 60 – 90, 55 – 86, 68 – 98 za svaki pojedinačno, što upućuje na činjenicu da su razlike u ocjenama ozbiljan problem, pogotovo kada se radi o ocjenjivanju na temelju kriterija vrednovanja koje su nastavnici sami izradili. Za razliku od toga, ocjene koje su isti ocjenjivači dali istim esejima koristeći se rubrikama u procesu vrednovanja nisu tako drastično varirale te postoji svojevrsna podudarnost kod ocjenjivača: ocjene za treći esej su 66 i 72, za peti esej 70 i 74, a za sedmi esej 88 i 90. Ocjene koje je pet ocjenjivača dalo na temelju kriterija vrednovanja koje su sami napravili i na temelju rubrika prikazane su na Slikama 1 i 2 kako bi se podudarnost jasnije uočila.

Prikaz 1

Prikaz 1 pokazuje distribuciju ocjena koje je svaki ocjenjivač dao esejima na temelju vlastitih kriterija vrednovanja. Može se vidjeti da je svaki ocjenjivač dao drugačiju ocjenu, što upućuje na neslaganje i ozbiljna odstupanja.

Međutim, analiza Prikaz 2, koja prikazuje ocjene dodijeljene s pomoću rubrika, otkriva da je distribucija manja, a uglavnom i jednaka. Kako se može vidjeti na slici, krivulje koje prikazuju ocjenjivače i njihove ocjene uglavnom se podudaraju, što znači da su ocjene bile slične te da je primjenom rubrika postignut konsenzus u ocjenama. Može se zaključiti da je ocjenjivanje eseja uz pomoć rubrika dovelo do veće podudarnosti u ocjenama.

Prikaz 2

Kendallov koeficijent konkordancije (Kendallov tau w koeficijent) koristio se kako bi se statistički odredila pouzdanost ocjenjivača u ocjenjivanju uz pomoć vlastitih kriterija vrednovanja i uz pomoć rubrika. Neparametrijska statistička analiza, Kendallov koeficijent konkordancije, koristi se za određivanje stupnja slaganja između više od dva ocjenjivača koji imaju svoje vlastite procjene o kvalitativnim kategorijama i za određivanje pouzdanosti i kompatibilnosti ocjenjivača (Akbulut, 2010, str. 174; Alpar, 2012, str. 464). Prema Kendallovu tau W, koeficijent je između 0 (nema podudarnosti) i +1 (potpuna kompatibilnost). Ako je vrijednost bliža broju 1 (Can, 2014, str. 376), upućuje na visok stupanj kompatibilnosti. U sklopu ovoga istraživanja analizirana je kompatibilnost između ocjenjivača, i to najprije za ocjene koje su dali u skladu s vlastitim kriterijima vrednovanja, a zatim u skladu s rubrikama (Tablica 4).

Tablica 4

Relevantni izračuni provedeni s vrijednostima u Tablici 4 pokazali su da je Kendallov tau (W) koeficijent iznosio 0,17 ($p>0,05$) za vrednovanje na temelju kriterija vrednovanja koje su izradili nastavnici. Kako je vrijednost blizu 0, ona upućuje na vrlo nizak stupanj kompatibilnosti u ocjenama koje su dali ocjenjivači. Međutim, Kendallov tau (W) koeficijent za vrednovanje na temelju rubrika iznosio je $W=0,903$ ($p<0,05$). Uzimajući u obzir da je vrijednost blizu broja 1, može se primijetiti da je stupanj kompatibilnosti u ocjenama koje su dali ocjenjivači na temelju rubrika prilično visok.

Dakle, statistički je dokazano da je međusobna pouzdanost ocjenjivača u vrednovanju pisanih uradaka na temelju rubrika puno veća nego kod vrednovanja na temelju kriterija koje su nastavnici sami izradili.

Kojim se kriterijima vrednovanja i načinima ocjenjivanja koriste nastavnici/ocjenjivači koji su sudjelovali u istraživanju?

Rezultati dobiveni putem polustrukturiranih individualnih intervjua provedenih u prvoj fazi istraživanja, a koji se odnose na treće pitanje istraživanja, prikazani su u Tablici 5, a zatim su detaljno objašnjeni u tekstu ispod tablice.

Tablica 5 pokazuje da su O1, O2, O3, O4 i O5 svaki pojedinačno uzimali četiri, šest, četiri, sedam i četiri kriterija kada su vrednovali eseje na temelju kriterija vrednovanja koje su sami izradili. Ipak, postoje tri kriterija koje su svi ocjenjivači uzimali u procesu vrednovanja: karakteristike teksta, gramatika i vokabular. Osim toga, postoji i još jedna važna razlika s obzirom na sadržaj kriterija. Na primjer O5 nije spomenuo točnost u pisanju, a gramatiku je podijelio u dvije komponente (točnost u pisanju gramatičkih struktura i točnost u pisanju riječi) te ocijenio točnost u pisanju unutar kriterija točnosti pisanja riječi. Kako se ponovno može vidjeti u Tablici 5, O1, O2, O3, O4 i O5 su svaki pojedinačno zanemarili tri, jedan, tri, jedan i dva kriterija tijekom vrednovanja eseja. O1 i O3 su zanemarili istu skupinu od tri kriterija (izvršenje zadatka, sintaksa, točnost u pisanju), O2 nije uzimao kriterij koherencija/kohezija, O4 zanemario je sintaksu, a O5 u procesu vrednovanja nije uzeo u obzir izvršenje zadataka.

Što se tiče zanemarenih kriterija, O1 je o izvršenju zadatka izjavio: „Što se mene tiče, taj je kriterij uključen u karakteristike teksta. Ne smatram ga zasebnim kriterijem." Što se tiče sintakse, isti je ocjenjivač rekao da ona pripada gramatici i kao takva se vrednuje. Kada se radi o točnosti u pisanju riječi, O1 je rekao da ono nema važnu ulogu i objasnio: „Ono što je bitno jest shvatiti što je student htio reći. Ako je razumljivo, pogreške u pisanju nisu tako važne." Dodao je: „Ako tekst nije razumljiv, to vrednujem u sklopu koherencije/kohezije, i ocijenim nižom ocjenom ako je potrebno." Čini se da je nemoguće složiti se s tim ocjenjivačem, jer je točnost u pisanju ključna komponenta dobro napisanoga eseja, a netočno napisane riječi znatno utječu na razumijevanje teksta, čine ga nejasnim i zamaraju čitatelje.

Tablica 5

Rezultati prvoga kruga polustrukturiranih intervjua

Ocjenjivači	Kriteriji vrednovanja koje su ocjenjivači uzeli u obzir	Kriteriji vrednovanja koje su ocjenjivači zanemarili	Sustav ocjenjivanja
O1	* Karakteristike teksta * Gramatika * Vokabular * Koherencija/kohezija	* Izvršenje zadatka * Sintaksa * Točnost u pisanju	* Karakteristike teksta: 50 bodova – Relevantnost argumenata – Odlomci – Ima li esej uvod, glavni dio i zaključak? *Gramatika: 15 bodova – (-1) bod za svaku pogrešku (1 bod se oduzima za pogreške koje se ponavljaju) * Vokabular: 15 bodova – Ukupna procjena * * Koherencija/kohezija: 20 bodova – Ukupna procjena
O2	* Izvršenje zadatka * Gramatika * Točnost u pisanju * Vokabular * Duljina teksta * Karakteristike teksta	*Koherencija/kohezija	* Izvršenje zadatka: 10 bodova * Duljina teksta: 5 bodova * Karakteristike teksta: 15 bodova – Argumenti – Teza/antiteza Ostalih 70 bodova: * Gramatika: (-½) boda za svaku pogrešku * Točnost u pisanju: (-½) boda za svaku pogrešku * Vokabular: (-½) boda za svaku pogrešku
O3	* Koherencija/kohezija * Vokabular * Gramatika * Karakteristike teksta	* Točnost u pisanju * Izvršenje zadatka * Sintaksa	* Sustav je binaran (ili 1 ili 0) 0 bodova za svaku pogrešku +1 bod za svaku točnu upotrebu * Vokabular – ukupna procjena (ne koristi se binarni sustav) * Opća procjena argumenata i karakteristika teksta (ne koristi se binarni sustav) * Ocjenjivanje: – 100/4= 25 bodova – Gramatika: 25 bodova – Koherencija/kohezija: 25 bodova – Vokabular: 25 bodova – Argumenti/Karakteristike teksta: 25 bodova

Ocjenjivači	Kriteriji vrednovanja koje su ocjenjivači uzeli u obzir	Kriteriji vrednovanja koje su ocjenjivači zanemarili	Sustav ocjenjivanja
O4	* Izvršenje zadatka * Vokabular * Gramatika * Točnost u pisanju * Koherencija/kohezija * Izražavanje mišljenja * Karakteristike teksta	* Sintaksa	* Izvršenje zadatka: 5 bodova * Vokabular: 10 bodova * Gramatika: 15 bodova * Točnost u pisanju: 15 bodova * Koherencija/kohezija: 20 bodova * Izražavanje mišljenja: 20 bodova * Karakteristike teksta: 15 bodova – Struktura teksta – Uvod? ➔ Ukupna procjena svega
O5	* Karakteristike teksta * Vokabular * Gramatika: točnost u pisanju gramatičkih struktura i točnost u pisanju riječi *Koherencija/kohezija	* Izvršenje zadatka * Sintaksa	* Ocjenjivanje: – 100/4 = 25 bodova – Karakteristike teksta: 25 bodova – Gramatika: 25 bodova – Koherencija/kohezija: 25 bodova – Vokabular: 25 bodova ➔ Ukupna procjena svega

O2 ovako objašnjava zašto u procesu vrednovanja eseja nije uzeo u obzir koherenciju/koheziju: „To se odnosi na uporabu veznika i mislim da to pripada gramatici. Zašto bih to trebao posebno vrednovati?“ Na pitanje: „Zašto niste točnost u pisanju riječi uzeli kao jedan od kriterija vrednovanja?“, O3 je odgovorio: „Točnost u pisanju riječi ne utječe na značenje. Ako postojeće pogreške u pisanju ne utječu na razumijevanje teksta, ne tretiram ih kao pogreške.“ Nadalje, O3 je rekao da je obratio pažnju i na sintaksu, no da ju je ocjenjivao u sklopu gramatičke komponente. Što se tiče izvršenja zadatka, rekao je:

> „Ne smatram da je potrebno zasebno ocjenjivati izvršenje zadatka. Od studenata se očekuje da će izvršiti zadatak i slijediti upute. Ukoliko to ne učine, znači da su skrenuli s teme, za što postoji utvrđena ocjena i nema potrebe za daljnjim ocjenjivanjem eseja. Najveća ocjena koju ću dati za takav esej je 5 od 100.“

Pri vrednovanju ukupno sedam kriterija, O4 je zanemario samo sintaksu, a uzeo ju je u obzir kao dio gramatike. Objašnjenje koje je O5 dao o tome zašto nije vrednovao izvršenje i sintaksu, bilo je da previše kriterija izaziva zbunjenost. Stoga je sintaksu smatrao dijelom gramatike, a izvršenje zadatka dijelom karakteristika teksta. Dakle, činjenica da ocjenjivači uzimaju u obzir različite kriterije tijekom procesa vrednovanja pisanoga uratka, a zanemaruju neka druga mjerila, razlog je niskoj individualnoj pouzdanosti ocjenjivača u vrednovanju eseja na temelju kriterija vrednovanja koje su izradili nastavnici.

Tablica 6 pokazuje korake u sustavu vrednovanja koje je svaki ocjenjivač opisao tijekom intervjua.

Tablica 6

Koraci u ocjenjivanju svakog ocjenjivača

	Ocjenjivač 1	Ocjenjivač 2	Ocjenjivač 3	Ocjenjivač 4	Ocjenjivač 5
1	Čitanje svakoga eseja i ispravljanje pogrešaka na svakome od njih (kao što su gramatičke pogreške), davanje pozitivne i negativne povratne informacije ("Dobro", "Dobro opažanje", "Jeste li sigurni?")	Čitanje svakoga eseja i ispravljanje gramatičkih pogrešaka i pogrešaka u pisanju, ocjenjivanje s +10 ili s -10 s obzirom na izvršenje zadatka i duljinu teksta.	Čitanje svakoga eseja i ispravljanje gramatičkih pogrešaka.	Čitanje svakoga eseja i ocjenjivanje s obzirom na izvršenje zadatka (5 bodova), karakteristike teksta (15 bodova) i izražavanje ideja (15 bodova).	Čitanje svakoga eseja i ocjenjivanje s obzirom na gramatiku (25 bodova) i vokabular (25 bodova).
2	Čitanje svakoga eseja i ocjenjivanje s brojem bodova do 50 s obzirom na karakteristike teksta.	Ponovno čitanje svakoga eseja i ocjenjivanje s obzirom na koherenciju/ koheziju (5 bodova) i argumente i uvjerljivost (15 bodova).	Ocjenjivanje argumenata s najviše 25 bodova, koherencije/ kohezije s najviše 25 bodova i vokabulara s najviše 25 bodova.	Provjera ispravaka u esejima i ocjenjivanje s maksimalno 15 bodova za točnost u pisanju i još maksimalno 15 bodova za gramatiku.	Ponovno čitanje svakoga eseja i ocjenjivanje s maksimalno 25 bodova za karakteristike teksta i maksimalno 25 bodova za koherenciju/ koheziju.
3	Provjeravanje gramatičkih pogrešaka i oduzimanje po 1 boda za svaku pogrešku (od maksimalno 15 bodova) (pogreške koje se ponavljaju uzimaju se u obzir samo jednom).	Ocjenjivanje svakoga eseja s maksimalno 70 bodova te oduzimanje po ½ boda za svaku gramatičku pogrešku, pogrešku u pisanju i za pogrešnu upotrebu riječi.	Ponovno čitanje svakoga eseja i provjera gramatike, za svaku pogrešku daje se 0 bodova, a 1 bod za svaku pravilnu upotrebu. Maksimalan broj bodova je 25.	Ocjenjivanje vokabulara s maksimalno 10 bodova i koherencije/ kohezije s maksimalno 20 bodova.	Zbrajanje svih bodova i ocjenjivanje svakoga eseja s ocjenom do 100.
4	Ocjenjivanje vokabulara s maksimalno 15 bodova.	Zbrajanje svih bodova i ocjenjivanje svakoga eseja ocjenom do 100.	Zbrajanje svih bodova i ocjenjivanje svakoga eseja ocjenom do 100.	Zbrajanje svih bodova i ocjenjivanje svakoga eseja ocjenom do 100.	
5	Ocjenjivanje koherencije/ kohezije s maksimalno 15 bodova.				
6	Zbrajanje svih bodova i ocjenjivanje svakoga eseja ocjenom do 100.				

Analiza sustava ocjenjivanja kojima su se koristili ocjenjivači pokazala je da svaki ocjenjivač ima svoj vlastiti način ocjenjivanja i da ne postoje razlike samo u komponentama koje su ocjenjivali, nego i da ocjenjivači iste komponente ocjenjuju na drugačiji način.

Što se tiče koraka u procesu ocjenjivanja eseja, O1 ih ima šest, O2, O3 i O4 imaju po 4 koraka, a O5 tri koraka. Na prvi pogled, usporedba koraka pokazuje da O5 ima

sustav ocjenjivanja koji oduzima najviše vremena i energije. Ipak, sustav ocjenjivanja O2, koji se sastoji od četiri koraka, tijekom kojih ocjenjivač oduzima po pola boda za svaku gramatičku pogrešku i pogrešku u pisanju i vokabularu, a zatim zbraja bodove (do maksimalno 70 bodova), ipak je najzamorniji i oduzima najviše vremena. Ocjenjivači O1 i O3 također su davali/oduzimali +/-1 bod za pogreške. Međutim, O4 i O5 su rekli da se nisu koristili tom tehnikom ocjenjivanja i da su eseje vrednovali u skladu s kriterijima i razinama znanja studenata, kao i usporedbom s esejima drugih studenata.

O4 ovako je opisao svoje ideje:

> „Definiram svoje kriterije i način ocjenjivanja u skladu s razinom znanja svojih studenata; ipak, ponekad usporedim eseje s onima drugih studenata i tada ocjenjujem određenu komponentu. Na primjer, ako student s B1 razinom znanja napiše esej s boljom gramatičkom točnošću nego njegovi kolege, onda mu dam više bodova nego ostalima. Smatram da on to zaslužuje; mislim da su i drugi trebali pokazati takvo znanje."

S obzirom na zajedničke kriterije koji se koriste tijekom vrednovanja ocjene variraju ovisno o ocjenjivaču, a važnost pojedinih kriterija također se razlikuje od ocjenjivača do ocjenjivača. Na primjer, „karakteristike teksta", kao zajednički kriterij kod svih ocjenjivača, može dobiti maksimalno 50 bodova kod O1, 15 kod O2 i O4, a 25 bodova kod O3 i O5. O1 smatra da su argumenti i uvjerljivost važniji od ostalih komponenti. Još jedan zajednički kriterij, gramatika, može dobiti maksimalno 15 bodova kod O1 i O4, a 25 bodova kod O3 i O5. Nije jasno koliko je bitnom tu komponentu smatrao O2, jer je ocjenjivao gramatiku, točnost pisanja i vokabular kao cjelinu, s najvećom ocjenom 70. Na kraju, vokabular, koji je još jedan zajednički kriterij svih ocjenjivača, kod O1 može dobiti maksimalno 15 bodova, kod O3 i O5 25, a kod O4 10 bodova. O3 i O5 ne smatraju da su jezične karakteristike jako važne.

Što ocjenjivači misle o upotrebi rubrika?

Rezultati su podijeljeni u tri kategorije, na temelju analize polustrukturiranih individualnih intervjua provedenih s pet nastavnika francuskoga jezika/ocjenjivača kako bi se ispitalo njihovo mišljenje o upotrebi rubrika odmah nakon ocjenjivanja 10 eseja sb pomoću tih rubrika. Tri kategorije rezultata bile su: korisnost rubrika u procesu vrednovanja; korisnost rubrika za studente i korisnost rubrika za nastavnike.

Svaka kategorija koja je nastala kao rezultat analize intervjua ispitana je u potkategorijama. Vrijednosti frekvencije prikazane u tablicama (Tablica 7, 8 i 9) koje pokazuju rezultate upućuju na to koliko su puta ocjenjivači spomenuli određenu situaciju/mišljenje. Zbog toga što je jedan sudionik nekoliko puta ponovio svoje ideje u vezi s različitim temama, ukupan broj frekvencija veći je od broja intervjuiranih sudionika.

Tablica 7

Kako je prikazano u Tablici 7, ocjenjivači su najčešće spominjali korisnost rubrika u procesu vrednovanja, i to uglavnom objektivnost. Svih pet ocjenjivača slaže se da su

kriteriji u rubrikama eliminirali osobne stavove i prosudbe. Evo što je svaki ocjenjivač rekao o toj korisnosti:

> „Obično me impresioniraju dobro razrađeni eseji. Kada pročitam dobar esej, ponekad se vratim na druge eseje koje sam već ispravio i dam im nižu ocjenu, jer smatram da su svi studenti na istoj razini znanja. Dakle, ako jedan od njih napiše kompleksnu rečenicu, i ostali bi to trebali također. Ako jedan od njih upotrebljava izraz iz višeg registra, i ostali studenti to moraju. Ali rubrike su to spriječile. Mislim da mogu svakoga studenta ocijeniti neovisno o ostalim studentima. Ukratko, bio sam poprilično objektivan u ocjenjivanju." (O4)

> „Danas sam nešto primijetio: prije sam obično ocjenjivao eseje na temelju vlastitih kriterija, vrijednosti i shvaćanja. Sada imam malo mjesta za osobne prosudbe." (O2)

> „Razlog zašto sam neke eseje prije ocjenjivao nižom ocjenom jest taj što sam gramatiku cijenio više nego bilo koju drugu komponentu. Ako je u eseju napravljeno puno gramatičkih pogrešaka, moj stav prema tom eseju bio je negativan… Mislim, osjećao sam određenu vrstu predrasude prema njemu. Na primjer, student pokušava objasniti neki svoj argument i napravi ozbiljnu gramatičku pogrešku. Ja se usredotočim samo na tu pogrešku, a sam sadržaj eseja uopće ne gledam ako ima puno gramatičkih pogrešaka. Ipak, primjena rubrika me je ovaj put zaustavila. Gramatičke pogreške nisu prekrile dobre strane eseja." (O3)

> „Primjena rubrika jako je objektivna. Bez obzira na to koliko se trudim profesionalno odraditi ocjenjivanje, uvijek imam osjećaj da je neka vrsta subjektivnosti stalno prisutna i da nešto nedostaje… no uz rubrike sve je bilo drugačije. Naravno, ne mogu reći da primjena rubrika eliminira sve probleme, pogotovo onda kada nemam volje ocjenjivati eseje. Ipak, one su koristan alat za provedbu objektivnijeg procesa vrednovanja." (O1)

Još jedna dobrobit primjene rubrika u procesu vrednovanja jest činjenica da one olakšavaju vrednovanje eseja i pružaju skup praktičnih kriterija. U tome se slažu svi ocjenjivači, a isto su spomenuli 15 puta tijekom intervjua. Ocjenjivači misle da rubrike uvelike smanjuju količinu vremena potrebnoga za ocjenjivanje eseja, a četvero njih (O1, O2, O3 i O5) spomenulo je njihovo lako tumačenje i primjenu, iako unutar rubrika ima puno kriterija. Ovo su bitni osvrti:

Smanjuje vrijeme potrebno za vrednovanje:

> „Uz rubrike sam imao priliku brzo i ugodno ocijeniti eseje. Možda mi je za prvi esej trebalo najviše vremena jer sam se prvi put koristio rubrikama, ali tada sam se naviknuo na njih i ostale sam radove brzo ocijenio." (O1)

> „Lakše je i brže raditi uz pomoć rubrika. Ocjenjivač bi trebao pažljivo pročitati i shvatiti rubrike prije nego što započne s ocjenjivanjem." (O2)

„U početku je primjena rubrika bila pomalo zamorna jer ih najprije trebate pažljivo pročitati i razumjeti. Ali poslije to ide puno brže. Trebalo mi je manje vremena za ocjenjivanje tih eseja. Istodobno štedimo vrijeme i provodimo uspješan proces vrednovanja." (O4)

„Štede vrijeme. Nakon prva dva eseja sve ide brže. Mogu reći da sam sada puno brži." (O5)

„Nije bitno koliko imam eseja. Sustav je pripremljen i ocjenjuješ prema njemu. Ocjenjivanje s pomoću rubrika puno je brže čak i kada ima više eseja." (O5)

„Osim toga, ne bismo trebali zaboraviti kako je praktično ocjenjivati esej s maksimalno 25 bodova, a onda ukupan zbroj pomnožiti s 4 kako bismo došli do maksimalnih 100 bodova. To olakšava posao i ubrzava ga. Također utječe i na smanjenje količine vremena potrebne za ocjenjivanje." (O2)

„Trebali smo ocijeniti eseje do maksimalne ocjene 100. Davanje ocjena do 25 i onda množenje rezultata s 4 olakšalo nam je proces. Nisam osjećao umor i osjećam da je finalna ocjena pokazala u kojoj je mjeri esej dobro napisan. Ozbiljno, proces je puno brži i lakši." (O3)

Lagana primjena:

„Bilo mi je lakše. Kriteriji unutar rubrika su izravni, jasni i relevantni. Bilo je jako jednostavno dati ocjenu u skladu sa skupinom kriterija. Mislim, kada razmislim o tome, ocijenio sam različite podvještine i zasebno sam ih vrednovao. Ocjenjivanje eseja nije bilo teško. Sve što sam trebao bilo je odabrati broj bodova za svaku komponentu." (O1)

„Znate da je ocjenjivanje eseja težak i zamoran posao, ali ovaj put nije bilo tako; bilo je vrlo lagano ocjenjivati eseje. Razlog je činjenica da je sve bilo kristalno jasno. Pročitao sam esej, razmislio o sadržaju, gramatičnosti i njegovim nedostacima putem primjene rubrika. Jako ih je lagano primjenjivati. Sve što trebate jest odabrati odgovarajući broj bodova za svaku komponentu; da biste to napravili, trebate provjeriti rubriku i vidjeti koliko je student dobar s obzirom na određeni kriterij." (O5)

Na kraju, ocjenjivači su također spomenuli valjanost kao još jednu dobrobit uvođenja rubrika u proces vrednovanja. Svi se ocjenjivači slažu da je s pomoću rubrika moguće procijeniti vještinu pisanja sigurno i na odgovarajući način te da su procjene o rezultatu studenata realistične, a zaključci o rezultatima odgovarajući. Ovo su bili dojmovi o toj potkategoriji:

„Nije se radilo samo o ocjenjivanju eseja, nego i o vrednovanju svih relevantnih kriterija kojima bi raspravljački esej trebao odgovarati. Drago mi je da ocjena koju sam dao odgovara kvaliteti sadržaja toga eseja… Ne postavljam si pitanja poput: „Je li ocjena previsoka li preniska? Je li vrednovanje bilo objektivno?"…

> Inače si postavljam ta pitanja… Često imam dojam da sam nešto zaboravio ili preskočio, npr. jesam li bio objektivan u ocjenjivanju svih eseja… to sam mislio… zahvalan sam na tim razrađenim kriterijima unutar rubrika." (O1)
>
> „Mogao sam ocjenjivati različite aspekte eseja. Ukupan broj bodova bio je dovoljno dobar da odgovara kvaliteti eseja. Mislim da sam vrednovao sve što sam trebao i da sam to napravio ispravno i na odgovarajući način. Vrednovanje je bilo relevantno, kao što bi i trebalo biti… Barem je to moj dojam." (O2)
>
> „Nikada prije nisam upotrebljavao tako sustavnu i organiziranu metodu pri ocjenjivanju pisanih uradaka. Što ocijeniti i koju najveću ocjenu dati potpuno je jasno s pomoću rubrika. […] Mogao sam procijeniti svaki kriterij bez usporedbe s drugima te smatram da sam zahvaljujući rubrikama bio jako dobar u vrednovanju eseja. Postoje kriteriji koji odgovaraju broju bodova koje dodijelim, znate na što mislim? …završna ocjena, rezultat, broj bodova, kako god ih hoćete nazvati, adekvatna je; postoje stvari, kriteriji, koji podupiru i objašnjavaju ocjenu koju dam." (O5)

Tablica 8

Kako pokazuje analiza polustrukturiranih intervjua, ocjenjivači su primijetili da bi rubrike također mogle biti korisne za studente, pogotovo u vezi s dva aspekta (Tablica 8): studenti mogu prepoznati područja koja trebaju poboljšati i prepoznati što se u pisanim zadatcima od njih očekuje. Svi ocjenjivači spomenuli su te dvije potkategorije ukupno 15 puta tijekom intervjua. Ocjenjivači ih nisu samo spomenuli, već su detaljno objasnili kako rubrike mogu biti korisne za studente s obzirom na te dvije potkategorije. Što se tiče ocjenjivača, studenti uz pomoć rubrika mogu sami provesti određenu vrstu samovrednovanja vlastitoga pisanog uratka i mogu naučiti koje su podvještine potrebne, kao i što bi trebali poboljšati kako bi mogli napisati kvalitetan pisani uradak. Ovo su njihova mišljenja o potkategoriji „Prepoznati područja koja je potrebno poboljšati":

> „Čini mi se da studenti mogu vidjeti koliko su pogrešaka učinili i kakve su te pogreške. Mogu procijeniti u kojim su kriterijima dobri, a u kojima nisu te mogu poboljšati slabije strane na sljedećem zadatku pisanja." (O3)
>
> „Mislim da rubrike sadrže više od samih bodova i ocjena. Čini se da se njima koristimo kako bismo došli do konačne ocjene, ali one također ističu i ono što studenti znaju i ne znaju. […] Što kada bih rubrike dao studentima kada im vratim ocijenjene eseje i rekao: „Vidite, ovo je vaš esej, a ocjena je 92. A ovo su kriteriji kojima sam se služio pri ocjenjivanju. Pogledajte ih." Tada bi student mogao jasno uočiti vlastite pogreške i shvatiti zašto je izgubio tih 8 bodova. Ne bih trebao ništa detaljno objašnjavati. Na primjer, student može reći: „Vokabular mi je slab." U skladu s tim bit će pažljiviji s vokabularom kada bude pisao sljedeći zadatak; pokušat će poboljšati slabije strane." (O4)

„Kada studentima damo rubrike, moći će analizirati svoje pogreške. Bit će dovoljno da detaljnije prouče rubriku. Također mogu usporediti svoj esej s rubrikom i sami provesti svojevrsnu analizu pogrešaka. Tijekom sljedećega zadatka pisanja bit će pažljiviji u slabijim područjima koje su utvrdili proučavanjem rubrika. Barem je to ono što bih ja učinio da sam student." (O5)

Ovo su mišljenja o potkategoriji „Prepoznati što se u pisanim zadatcima očekuje od studenata":

„Studenti će moći uvidjeti što se od njih očekuje i što vještina pisanja zahtijeva ako im uz ispravljene eseje damo i rubrike. Ipak, što se mene tiče… Studenti bi trebali dobiti rubrike na početku semestra i analizirati ih. Tako bi zaista razumjeli što se od njih očekuje… To bi im zasigurno pomoglo… Dao bih im rubrike na samome početku, da se mene pita. Čak bih im rekao da ih zadrže i njima se koriste pri pisanju domaće zadaće." (O1)

„Stvarno, studenti mogu razumjeti očekivanja, sve što se od njih u zadatku pisanja očekuje i sastavnice pisanoga uratka." (O3)

„Studenti će moći vidjeti što trebaju učiniti kako bi napisali dobar esej te što su sastavnice dobroga eseja. Bit će svjesni svega što trebaju napraviti kako bi bili uspješni. Sve će im biti jasno kada vide popis kriterija. [O4 postavlja pitanje istraživaču: Želim nešto pitati. Je li moguće prilagoditi ove rubrike svim razinama? Razini A2, na primjer? Istraživač: Da.] Kako budu postajali sve vještiji, vidjet će da se očekivanja također mijenjaju. Mislim, ako na B2 razini ocjenjujemo neke kriterije višom ocjenom, tada će studenti to primijetiti i misliti da su ti kriteriji važniji. Mogu shvatiti i sami: „Broj bodova za ovaj kriterij je sada viši, pa bih u tome dijelu trebao biti bolji." (O4)

Tablica 9

Ocjenjivači su također naveli da su rubrike korisne i njima. Primijetili su da su studentima mogli dati detaljnu povratnu informaciju o esejima s pomoću rubrika te da se njima također mogu koristiti kada studente poučavaju vještini pisanja (Tablica 9). Što se tiče tih dviju potkategorija, četvero ocjenjivača (O1, O2, O4 i O5) složilo se s prvom (O1, O2, O3), a njih troje istaknulo je drugu potkategoriju. Ovo su mišljenja o potkategorijama koje su ocjenjivači spomenuli sedam puta:

Detaljna/učinkovita povratna informacija:

„Da budem iskren, ponekad mi je teško odgovoriti kada me studenti pitaju o pogreškama u esejima, jer zaboravim njihov sadržaj. Zato mislim da će i meni rubrike dobro doći. Kada ih pogledam, mogu lako reći da je student X dobar ili slabiji u određenome području. Smatram da rubrike pružaju bolja objašnjenja. Osim toga, ne mogu dati dobru povratnu informaciju kada se radi o velikom broju eseja. Na početku mogu dati bolju povratnu informaciju

[…] Studenti neće morati trčati za nama kako bi uvidjeli svoje pogreške. Sada količina napisane povratne informacije na eseju nije toliko bitna jer su rubrike dovoljne; one su također vrsta povratne informacije, i to bolje." (O2)

„Mogu lako objasniti zašto sam studentu X dao određen broj bodova kada se služim rubrikama, a to mogu napraviti i sa svim esejima jer se vrednuju s pomoću istih kriterija i pod istim uvjetima. Hm… Recimo to ovako: ponekad, kada studentima vratim ocijenjene eseje i kada me pitaju zašto su dobili slabu ocjenu, moram ponovno pogledati određeni esej da bih ga se prisjetio te onda mogu reći nešto o tome… Razlog zašto to činim jest taj što ponekad ne oduzmem bodove za istu pogrešku kod drugih studenata; mislim, nekima oduzmem bodove, a nekima ne… Kada vratim studentima ocijenjene eseje da prouče svoje pogreške, kažem im da analiziraju samo svoj esej. Kažem im da gledaju samo svoj esej i da me pitaju ako imaju pitanja bez prethodnog razgovora s kolegama. Ovako će rubrike stati tome na kraj… Mogu jednostavno objasniti što nije u redu u njihovim esejima i objasniti što se uzima u obzir u vrednovanju. Mogu dati dosljedno i razumno objašnjenje svakome studentu, ne moram ponovno pregledati esej – bit će dovoljno pogledati rubrike." (O4)

„Obično na eseju ispravim 2 ili 3 pogreške, a studenti me pitaju jesu li te pogreške razlog lošim ocjenama. Nije lako kontinuirano davati povratnu informaciju. Međutim, kada im budem dala rubrike, svaki će student imati svoju vlastitu povratnu informaciju." (O5)

Primjena u nastavi:

„Ako svi zajedno proučimo rubrike… mislim, zajedno sa studentima, u učionici i putem interakcije… Ako svi zajedno u učionici razjasnimo što znači kriterij X i što mjeri kriterij Y, razgovaramo i o vještini pisanja i tako je obrađujemo... Ako provedemo kratke aktivnosti vezane uz svaki pojedinačni kriterij, […] može se organizirati uspješna vježba za kriterij „koherencija/kohezija". Mislim da se svi kriteriji unutar rubrike mogu pojedinačno analizirati: možemo studente poučiti što to znači uspješno pisati koristeći se rubrikama kao nastavnim materijalom… Na primjer, što je to pisanje? To je ono što podrazumijeva zadatak pisanja… to je vještina koja se sastoji od drugih vještina navedenih u rubrici. Možemo tvrditi da raspravljački esej uključuje kriterije definirane rubrikama te proučiti rubrike." (O1)

„Kriteriji unutar rubrika zapravo su podvještine potrebne za vještinu pisanja… Ako analiziramo svaki pojedinačni kriterij zasebno i objasnimo ga… razgovaramo o njihovoj funkciji i važnosti… ujedno provodimo i nastavu u kojoj poučavamo pisanje, tj. razgovaramo o teoriji pisanja… Teorijski, možemo razgovarati o tome što je neophodno za uspješan zadatak pisanja, o dijelovima pisanoga teksta i možemo vježbati. Na kraju, pisanje jest vještina, ali ona se

temelji na relevantnoj teorijskoj osnovi… Inače se taj teorijski dio u poučavanju vještini pisanja zanemaruje, a studenti odmah započinju vježbati pisanje. Međutim, ove rubrike omogućuju također i obradu teorije pisanja." (O2)

Rasprava

Postoje brojne varijable koje stoje na putu učinkovitom vrednovanju rezultata (Black, 1998), a varijabla ocjenjivača najčešće je istaknuta kada se radi o vrednovanju pisanih uradaka (Moskal, 2000). Pouzdanost i valjanost prosudbe ocjenjivača često se dovode u pitanje. S tim u vezi literatura navodi brojna istraživanja koja predlažu upotrebu rubrika, što je također i središnja tema ovoga istraživanja, kako bi se proveo pozitivan proces vrednovanja i kako bi se došlo do boljih rezultata.

Cilj je ovoga istraživanja bio odrediti pouzdanost (individualnu i međusobnu pouzdanost ocjenjivača) rubrika nakon procesa vrednovanja. Petero nastavnika francuskoga jezika koji nikada prije nisu upotrebljavali rubrike i koji su se uvijek koristili svojim vlastitim kriterijima, ocijenilo je 10 raspravljačkih eseja najprije s pomoću kriterija koje su sami izradili, a zatim uz pomoć rubrika, i to nakon tri mjeseca. Polustrukturirani intervjui pokazali su da su kriteriji varirali od ocjenjivača do ocjenjivača te je ispitana kvaliteta rubrika.

U skladu s prvim pitanjem istraživanja uspoređene su ocjene kojima su ocjenjivači ocijenili eseje, i na temelju vlastitih kriterija, i na temelju rubrika. Uočena je ozbiljna razlika u ocjenama. Utvrđeno je da neki ocjenjivači nisu istom ocjenom ocijenili esej primjenom dviju tehnika ocjenjivanja te da je razlika u ocjenama nekih eseja bila +/- 16. To znači da je individualna pouzdanost ocjenjivača, tj. podudarnost dviju ocjena koje je isti ocjenjivač dao istome eseju bila prilično niska.

Drugo pitanje istraživanja ispitivalo je međusobnu pouzdanost ocjenjivača usporedbom dviju ocjena koje su različiti ocjenjivači dali istome eseju primjenom različitih tehnika ocjenjivanja. Također je analiziran konsenzus među ocjenjivačima, tj. podudarnost ocjena koje su dali. Rezultati istraživanja pokazuju da je međusobna pouzdanost ocjenjivača bila visoka kada su se u procesu vrednovanja koristile rubrike. Uistinu, ocjene koje su ocjenjivači dali primjenom rubrika bile su vrlo slične, što pokazuje konsenzus među njima. Što se tiče ocjenjivanja uz pomoć kriterija koje su nastavnici sami izradili, uočeno je neslaganje među ocjenjivačima: dali su znatno drugačije ocjene istim esejima kada su se koristili kriterijima koje su sami izradili.

Rezultate koji su dobiveni u vezi s prvim i drugim pitanjem istraživanja potvrdila je statistička analiza, kao što je Spearmanov koeficijent korelacije ranga i Kendallov tau (W). Uočeno je da su ocjenjivači dosljedniji i pouzdaniji kada se koriste rubrikama nego kada se koriste vlastitim kriterijima. Ti rezultati imaju smisla jer ocjenjivačima rubrike služe kao vodič za ocjenjivanje i sadrže sve relevantne kriterije. Također opisuju različite razine kvalitete svakoga kriterija, što ujedno i naglašava kako bi ocjenjivači trebali ocjenjivati glavnu misao i sadržaj pisanoga uratka te o kojim bi kriterijima trebali posebno voditi računa (Arter i McTighe, 2001; McMillan, 2004).

Pregled literature obiluje istraživanjima u kojima se navodi da je ocjenjivanje uz pomoć rubrika dosljednije i pouzdanije, da ocjenjivačima pomaže u smanjenju međusobnih razlika (međusobna pouzdanost ocjenjivača) te da mogu smanjiti nedosljednosti u ocjenjivanju koje se javljaju zbog unutarnjih čimbenika (individualna pouzdanost ocjenjivača) (Hansson, Svensson, Strandberg, Troein, i Beckman, 2014; Jonsson i Svingby, 2007; Moskal i Leydens, 2000).

U skladu s kvalitativnim podacima prikupljenima tijekom polustrukturiranih intervjua koji su se provodili u vezi s trećim pitanjem istraživanja, vlastiti kriteriji kojima se ocjenjivači koriste kada ocjenjuju eseje značajno variraju, a broj se kriterija koji se koristi u vrednovanju značajno mijenja: neki ocjenjivači uzimaju u obzir samo četiri kriterija, a drugi šest ili sedam kriterija. Iako su tri kriterija (karakteristike teksta, gramatika i točnost u pisanju) zajednička svim ocjenjivačima, komponente koje se ocjenjuju unutar tih kriterija razlikuju se među ocjenjivačima, što znači da se ne slažu oko određenih mjerila, iako se generalno slažu s kriterijima. Osim toga, broj koraka u kojima čitaju i putem kojih ocjenjuju eseje također nije isti: dok O1 treba puno više vremena da bi ocijenio jedan esej zato što njegov proces ima šest koraka, O5 ocijeni esej u kraćem vremenu i u ocjenjivanju prolazi samo dva koraka. Sve to može dovoljno dobro objasniti što sve utječe na međusobnu pouzdanost ocjenjivača te zašto je ona tako niska kada se u procesu vrednovanja koriste kriteriji koje su nastavnici sami izradili. Različitost kriterija na kojima ocjenjivači temelje svoju ocjenu i različit broj bodova koje daju za određene kriterije utječe na konačnu ocjenu, što dovodi do različitih ocjena kojima su isti eseji ocijenjeni. McNamara (1996, str. 117) smatra da „vrednovanje rezultata zasigurno uključuje subjektivnu procjenu" i da ta subjektivnost često utječe na razlike u vrsti kriterija i postupaka ocjenjivanja ili u tumačenju kriterija ocjenjivanja. Barkaoui (2007, str. 86) je u svojem istraživanju pokazao da „su ocjenjivači glavni izvor različitosti kada se radi o ocjenama i donošenju odluka". Isto tako, Schoonen (2005, str. 1) je pokazao da „generaliziranje ocjena pisanih uradaka i utjecaj ocjenjivača i tema u velikoj mjeri ovise o načinu na koji se eseji ocjenjuju i o karakteristikama koje se ocjenjuju."

Na kraju, četvrto pitanje istraživanja imalo je za cilj utvrditi što nastavnici/ocjenjivači koji su sudjelovali u istraživanju i koji su prvi put upotrebljavali rubrike u ocjenjivanju eseja misle o rubrikama te je proveden drugi krug intervjua. Rezultati su pokazali da ocjenjivači posebno ističu da upotreba rubrika vodi objektivnijim ocjenama i da primjena vlastitih kriterija u ocjenjivanju vodi subjektivnim ishodima, usredotočeno je na manji broj kriterija i podrazumijeva gramatiku kao vodeći kriterij. Kako su ocjenjivači naveli, rubrike su ih oslobodile takvih pogrešaka koje su vodile subjektivnoj procjeni tijekom ocjenjivanja, jer im je taj alat ocjenjivanja pomogao da se usredotoče na kriterije uzimajući sve kriterije kao cjelinu: „Kada sam pred sobom imao mjerila za sve kriterije, nije mi bilo teško usredotočiti se. Koncentrirao sam se na svaki kriterij i radio u skladu s njim" (O2). Ahoniemi i Reinikainen (2006, str. 139) također su toga mišljenja: „…jedini način na koji se može postići objektivnost jest podijeliti ocjenjivanje na manje dijelove uz pomoć rubrika."

Što se tiče rubrika, „njihova praktičnost odnosi se na laganu primjenu" (Arter i McTighe, 2001, str. 49). Ocjenjivači koji su sudjelovali u ocjenjivanju tom pitanju pristupaju iz dva različita kuta. Prvo, eseji su bili brže ocijenjeni, jer su rubrike prikazale sve kriterije zajedno. Međutim, također su naglasili da su prije samoga ocjenjivanja morali dobro razumjeti rubrike i detaljno proučiti svaki kriterij. Neki su od njih izjavili da im je ocjenjivanje prvoga eseja išlo jako sporo, jer su istodobno pokušavali shvatiti rubrike, ali je proces išao sve brže i brže što su ocjenjivali veći broj eseja. K tome, ocjenjivači su spomenuli da je uz pomoć rubrika bilo lako ocijeniti eseje ocjenom do 100, što je također ubrzalo proces ocjenjivanja. Što se tiče O2, koji oduzima po pola boda za svaku pogrešku iz područja gramatike, točnosti pisanja i upotrebe riječi kako bi došao do ukupnoga broja bodova od 70 kada se koristi vlastitim kriterijima ocjenjivanja, možemo uočiti koliko su rubrike olakšale proces ocjenjivanja. Slično tome, rubrike su također korisne za ocjenjivače poput O1, koji od ukupno 15 bodova oduzima po 1 bod za svaku gramatičku pogrešku i za O3 koji daje 0 bodova za svaku pogrešku i 1 bod za svaku točnu upotrebu kako bi u gramatičkom dijelu došao do broja bodova od 25. Drugi razlog zašto je ocjenjivačima ocjenjivanje uz pomoć rubrika lakše i praktičnije jest taj što su rubrike olakšale proces ocjenjivanja, a nisu zasjenile sadržaj eseja: bili su zadovoljni što su precizno i lako mogli ocijeniti svaku podvještinu stavljanjem kvačica u pravokutnike. Bainer i Prter (1992, str. 12) su u svojem istraživanju objasnili da su se nastavnici složili da je upotreba rubrika „u vrednovanju eseja laka jer rubrike sadrže jasne smjernice koje trebaju slijediti te na taj način pružaju *osnovu za početak*."

Druga bitna činjenica koju su ocjenjivači spomenuli tijekom intervjua odnosi se na valjanost procesa vrednovanja. Kako navodi Nitko (2004, str. 34): „Kako bi potvrdili [svoju] interpretaciju i primjenu rezultata vrednovanja učeničkih radova, [ocjenjivači] moraju dati dovoljno dokaza da su interpretacija i primjena rezultata prikladni." To je upravo ono što su ocjenjivači spomenuli: primijetili su jasnoću kriterija i stupnjevito bodovanje unutar svakoga kriterija i izjavili kako im je to pomoglo da lako objasne zašto su određenom ocjenom ocijenili esej. Na primjer, O4 je izjavio sljedeće: „Pogledao sam kriterije i tada ocijenio esej. Čini mi se da moja procjena kvalitete eseja ima čvrsto uporište… Ukupna je ocjena održavala koliko je rad bio dobar i imao sam osjećaj da je moja procjena bila odgovarajuća i valjana." Isto tako, Bresciani i sur. (2004, str. 30) naveli su da rubrike „sprječavaju optužbe da ocjenjivači ne znaju ni sami što traže".

Na kraju, ocjenjivači su primijetili da bi rubrike bile korisne i studentima i njima samima. Kako navode, rubrike mogu pomoći studentima da prepoznaju područja koja moraju poboljšati te što se od njih u zadatcima pisanja očekuje, a nastavnicima mogu pomoći da daju detaljnu povratnu informaciju i pouče studente kako pisati. Zapravo, komentari ocjenjivača da bi upotreba rubrika bila korisna i za njih i za studente u skladu je s rezultatima mnogih istraživanja. Na primjer, referirajući se na brojne studije, Reddy i Andrade (2010, str. 437) su izjavili: „Kada se njima koriste učenici

kao dio formativnog ocjenjivanja radova na kojima rade, rubrike mogu i poučavati i ocjenjivati". Stevens i Levi (2013) smatraju da je davanje rubrika učenicima vrlo učinkovita strategija jer tako znaju što se od njih očekuje, a mnogi se učenici opuštenije osjećaju kada unaprijed znaju kriterije. Oakleaf (2019, str. 969) smatra da „rubrike pomažu učenicima da razumiju očekivanja nastavnika". Slično tome, Bresciani i sur. (2004, str. 30) kažu da rubrike „čine poznatima ključne kriterije kojima se učenici mogu koristiti u pisanju, pregledavanju i procjeni vlastitoga rada". Jaidev (2011, str. 7) je također naglasio da rubrike imaju ključnu ulogu u poboljšanju vještine pisanja, a učenici mogu bolje izraziti vlastito mišljenje zahvaljujući rubrikama: „Znanje o rubrikama koje se odnose na pisanje pomaže učenicima da budu odgovorniji u zadatcima pisanja i u njima stvara osjećaj svojevrsnoga vlasništva nad onime što su napisali". No, zbog toga što „rubrike nisu potpuno razumljive same po sebi" (Andrade, 2005, str. 29), možda nije dovoljno samo ih uručiti učenicima i reći da se njima koriste. „Učenicima je potrebna pomoć u razumijevanju rubrika i njihovoj primjeni" (Andrade, 2005, str. 29), upravo ih zato nastavnici trebaju objasniti učenicima, i to svaki kriterij pojedinačno. Ocjenjivači koji su sudjelovali u istraživanju, naglasili su da bi proučavanje rubrika zajedno sa studentima bio način učenja o vještini pisanja i ujedno i alat u nastavi u kojoj se obrađuje pisanje. Kako navode sudionici, razgovor o kriterijima i njihovo objašnjavanje u rubrikama pomoći će studentima da ili nauče ili zadrže podvještine pisanja. Arter i McTighe (2001, str. 10) su izjavili: „Jasno definirani kriteriji i vodič za ocjenjivanje više su od alata za evaluaciju na kraju nastavnog procesa – oni pomažu razjasniti nastavne ciljeve i služe kao ciljevi u nastavi". Što se tiče ocjenjivača, mogućnost davanja detaljnije povratne informacije još je jedna dobrobit upotrebe rubrika u procesu vrednovanja. Primijetili su da su na esejima napravili samo nekoliko ispravaka jer je proces davanja cjelokupne povratne informacije jako zamoran i oduzima puno vremena te da im je bilo teško objasniti studentima koje su pogreške napravili i zašto su dobili tu određenu ocjenu. Bilo im je drago što su rubrike bile korisne u pružanju detaljne, smislene i učinkovite povratne informacije jer su trebali samo u njih pogledati i pomoći studentima da shvate što trebaju popraviti u zadatcima pisanja. Istraživanja koja su proveli Stevens i Levi (2005), Reddy i Andrade (2010) i Brookhart (2013) također idu u prilog ovim spoznajama.

Zaključak

Kao zaključak možemo reći da se primjenom rubrika u procesu vrednovanja dolazi do pouzdanijih i dosljednih ishoda, kako pokazuju i kvantitativni i kvalitativni rezultati ovoga istraživanja. Rubrike su vjerodostojne i pouzdane, jer pomažu ocjenjivaču da zadrži postojane procjene različitih eseja. Stoga će vrednovanje pisanih uradaka s pomoću alata koji sadrže određene kriterije vjerojatno eliminirati nedosljednosti među ocjenjivačima. Na početku samoga procesa vrednovanja ocjenjivači obično određuju stupanj strogoće ili popustljivosti, a rubrike, za razliku od toga, zahtijevaju objektivnost jer kriteriji navedeni u njima odražavaju nastavne ciljeve i dostignuća

koja treba ocijeniti, sprečavaju dodavanje novih kriterija i sustavno se usredotočuju na iste komponente (Berthiaume i Collet, 2013). Naravno da nije moguće postići potpunu objektivnost zbog same činjenice da su ocjenjivači ljudi. Zato će uvijek postojati mogućnost subjektivnosti u ocjenjivanju. Ipak, rubrike su jedan od vrijednih alata s pomoću kojih se može smanjiti subjektivnost u vrednovanju učeničkih rezultata. Drugi čimbenik zbog kojega je upotreba rubrika korisna jest da se studente ne shvaća kao pasivne objekte ocjenjivanja, nego su i oni sami aktivno uključeni u proces vrednovanja. Činjenica da su ocjenjivači primijetili da bi studenti trebali zajedno s nastavnicima proučiti rubrike, može se uzeti kao pokazatelj toga. Na kraju, utvrđeno je da ocjenjivači koji su sudjelovali u istraživanju nisu prije toga imali iskustva s upotrebom rubrika. Zato je jako važno u nastavu na učiteljskim fakultetima uključiti edukaciju o upotrebi rubrika, kako nastavnici ne bi samo ocjenjivali eseje, nego da studentima mogu dati korisnu povratnu informaciju i riješiti probleme koji se javljaju kod vještine pisanja.

Dodatak

Rubrika za raspravljački esej – B1 razina – 25 bodova

Praćenje uputa									
Može prilagoditi svoju vještinu pisanja zadanoj situaciji. Može slijediti zadane upute u vezi s najmanjom dopuštenom duljinom teksta.	0	0,5	1	1,5	2				
Sposobnost izlaganja činjenica									
Može opisati činjenice, događaje i iskustva.	0	0,5	1	1,	2	2,5	3	3,5	4
Sposobnost izražavanja misli									
Može izložiti svoje ideje, osjećaje i/ili reakcije i izraziti svoje mišljenje.	0	0,5	1	1,5	2	2,5	3	3,5	4
Koherencija i kohezija									
Može povezati niz kratkih, jednostavnih elemenata u diskursu koji je tečan.	0	0,5	1	1,5	2	2,5	3		

Leksička kompetencija / Leksička točnost u pisanju

Raspon vokabulara Ima vokabular dostatan za pisanje o aktualnim temama; može parafrazirati ako je potrebno.	0	0,5	1	1,5	2
Kontrolirani vokabular Pokazuje dobru kontrolu nad osnovnim vokabularom, no veće pogreške se još uvijek javljaju kada izražava složenije misli.	0	0,5	1	1,5	2
Pravopis Leksička točnost u pisanju, interpunkcija i format su uglavnom dovoljno točni da se esej može lako čitati.	0	0,5	1	1,5	2

Gramatička kompetencija/ Gramatička točnost u pisanju

Razina razrađenosti rečenične strukture Dobra razrada jednostavnih rečeničnih struktura i učestalijih složenih rečeničnih struktura.	0	0,5	1	1,5	2
Izbor glagolskih vremena i načina Pokazuje dobro znanje iako se može primijetiti utjecaj materinskog jezika.	0	0,5	1	1,5	2
Morfosintaksa – Gramatička točnost u pisanju Slaganje u rodu, broju, zamjenicama, glagolskim nastavcima itd.	0	0,5	1	1,5	2