ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 16, No. 1, pp. 13-22, January 2025
DOI: https://doi.org/10.17507/jltr.1601.02

The Criteria and Challenges in Evaluating the Performance of the Writing Skills of German Language Students

Sadije Rexhepi
Faculty of Philology, University of Prishtina, Prishtinë, Kosovo

Çliresa Suka*
Faculty of Philology, University of Prishtina, Prishtinë, Kosovo

***Abstract*—The goal of this study was to learn how the performance of German as a Foreign Language (GFL) writing skills is evaluated, what criteria it is based on, and what challenges can be expected during the evaluation. Performance measurement is a challenge for many, and previous research and experience show that when it comes to productive abilities, writing takes a back seat in the classroom, where speaking skills are prioritized. Nonetheless, since the focus of this study is on German writing skills, it is important to clarify their role in GFL teaching. Although there are different aspects and criteria for the assessment of productive and writing skills, some are more important than others. The study takes a theoretical and empirical approach, utilizing a questionnaire and interviews to analyze the perspectives of both examinees (students) and examiners (teachers). For a justified assessment, certain quality criteria are considered, including objectivity, reliability, and validity. Additionally, the study also references Goethe, ÖSD, and TELC model tests, as well as other test forms used in private schools and university courses in Kosovo. Thus, this research focuses not only on the various forms and functions of performance measurement but also on the potential obstacles that may develop during assessment and how to avoid and fix these issues.**

***Index Terms*—writing skill, performance measurement, criteria of quality, content design, grammar**

I. Introduction

Studies on writing assessment in foreign or second languages have taken a range of approaches. Many authors, such as Storch (1999), Rösler (2012), and Decke-Cornill and Küster (2015), have investigated the writing assessment of German as a Foreign Language (GFL), whereas Kic-Drgas (2022) and Kast (2003) have studied writing assessment and its difficulties in English and German from various perspectives. Still others like Grotjahn and Kleppin (2017), Hinger and Stadler (2018), Glaboniat (2017), Bolton (1996), Kleppin (1998), and Albers et al. (1995) have explored the subject even deeper, specifically studying testing and assessment of writing skills in English and German, while Liu (2022) went so far as to analyze the reliability and feasibility of the integration of automated writing evaluation (AWE) and human evaluation in an English writing course.

Perhaps one of the most difficult tasks in the world of teaching and learning is performance measurement. As a result, one must first understand the term "performance measurement" and all its related terms so that a deeper dive into the topic is possible. Moreover, this research article will attempt to find out how the performance of the skill of writing in German is assessed, what difficulties arise in the process, and what criteria this assessment is based on. This will be accomplished by answering the following research questions:

1. How can performance in the open tasks, especially in writing skills, be measured or assessed as correctly as possible?
2. What difficulties are encountered when evaluating or correcting the open (writing) tasks?
3. Are there differences or similarities between the assessment criteria at examination centers (Goethe, TELC, and ÖSD), in private schools, or in university courses in Kosovo?

The assessment of productive skills is challenging for teachers as it is largely based on open-ended tasks. While evaluating these skills, and particularly writing abilities, a number of factors and standards must be considered, some of which are more significant than others. The hypothesis of this article is that content is more important for writing skills than other testable aspects.

II. Methodology

* Corresponding Author.

Various methods of data collection were used in this study, and the work takes both a theoretical and an empirical approach. Furthermore, a literature review was carried out for the theoretical data collection. This was important for the study in order to create a basis on which to build the practical or empirical part. Additionally, because this study is concerned with performance measurement, both the perspective of the examinees (learners and students) and the perspective of the examiners (teachers and lecturers) were considered. Moreover, after being fully informed about the project, the interview subjects gave their consent to participate and their permission for their interviews to be used for future work.

The interviews were conducted with lecturers from universities in Kosovo who teach and test German as a foreign language (GFL) to students at the B2 and C1 (Upper-Intermediate and Advanced, respectively) levels. Interviews were also conducted with teachers who teach and test German as a foreign language at B2 and C1 levels in courses at private schools. The aim of using interviews as a qualitative method was to find out how individual examiners measure performance and how they deal with the difficulties and problems of performance measurement. The interviews consisted of fifteen questions. The questions covered performance assessment in general and asked participants to identify the purposes and contributing components of performance measurement. Additionally, there were also questions that asked about how realizable and realistic the quality criteria for performance measurement are, to what extent formal and informal performance measurement can be used in different situations, and which factors play a fundamental role in this. It is worth noting that some of the interviewees worked as examiners in language centers as well as in universities, private schools, and private tutoring. It was found that assessment in language centers (Goethe, TELC, and ÖSD) is standardized and based on strictly defined criteria. These examiners work according to a set scheme and criteria. For this reason, the examiners were interviewed about their work in other areas (for example, as teachers at universities, in private language courses, or in schools).

The questionnaire had a standardized design and was distributed to students who are studying or have studied German at the B2 or C1 levels at two universities in Kosovo. Additionally, learners who are studying or have studied German at B2 or C1 levels at private schools or in private courses and examinees who have taken an exam at B2 or C1 levels at language centers also participated in the study. The questionnaire was created and distributed online using Google Forms. The questions in the questionnaire are similar to the questions in the interview and relate to the examinees' experiences with performance measurement. The other questions in the questionnaire deal with the forms of performance measurement, how often the respondents write in class (and for what purpose), and which skill is of greater importance to them. Other important questions of the study asked whether the criterion of objectivity is met to a greater extent in written examinations than it is in oral examinations. Questions were also asked about the assessment criteria and the forms of correction used during correction. The questionnaire was completed by one hundred respondents, and the data was analyzed quantitatively. A total of 19 questions were asked. The results of the interviews and the questionnaire are presented as the empirical part together with the theoretical part. In sum, the methodology of the thesis consists of a literature review, a quantitative analysis based on the questionnaire, and a qualitative analysis based on the interviews.

III. Performance Measurement

A. Explanation of Terms

Performance measurement is the result of measuring something that has been created or accomplished. As performance measurement encompasses various areas, the term can only be defined in the context of tests, assessments, examinations, etc. The related science of performance measurement is examination didactics. Tinnenfeld (2013) defines examination didactics as “the teaching of the examiner, i.e., the *what* and *how* of testing on a theoretical basis with an orientation towards practice and the respective target group” (p. 82).

According to the PONS dictionary (2008), “to perform” means “somebody accomplishes something (≈ accomplish) or somebody does or creates something that requires a lot of work” (p. 98). A similar definition can be found in the Langenscheidt dictionary (2008), where it means "to do or accomplish something that usually requires a great deal of effort” (p. 610). According to the Dudenredaktion (2015), the term measurement” is defined as "1. to carry out the measurement; 2. result of a measurement" (p. 1192). So according to the definitions, “performance measurement"’ refers to the result of a created or accomplished performance.

The definition of performance measurement, according to the authors Decke-Cornill and Küster (2015), is as follows:

> The verbs 'test' and 'measure' are generally used in the context of quantitatively recordable processes. 'Evaluate', on the other hand, refers to the evaluation of quantitative measurement results on the one hand but also to processes of qualitative judgment on the other. The latter aspect essentially coincides with the semantic field of 'evaluate'. All four have a moment of comparison. They either compare the relationship of what has been determined to a norm or an average, or they describe a development by recording a before and after state. (p. 260)

In the following chapters and subchapters, the related terms and aspects of performance measurement are explained and discussed in more detail.

B. The Functions of Performance Measurement

According to Grotjahn and Kleppin (2017), the functions of the various forms of performance measurement in the classroom include the diagnosis of learners' strengths and weaknesses, the promotion of support needs, the categorization of performance levels, the recognition of progress, and so on. For the learners or examinees, these forms of performance measurement have the function of providing information about their performance level, recognizing their own learning progress as a form of self-assessment, enabling them to obtain a certificate such as a Goethe or a TestDaF, or even get a job or be accepted by a university. The authors Hinger and Stadler (2018) define the function of performance measurement and testing in a similar way.

Performance measurement is an essential part of teaching in both educational institutions and private language courses. When it comes to language acquisition, such forms of performance measurement carried out on a specific day at a specific location are known as language proficiency tests. The aim of these tests is to check whether the learning objectives have been achieved.

As explained in the previous chapter on the methodology used in this study, interviews were conducted with teachers from different educational institutions, including universities, schools, and private language courses. Due to the differences and similarities between the educational institutions, the interviewees were categorized into three groups in order to compare their responses. The interviewees were then asked about the function of performance measurement.

According to the lecturers ("Dozenten" D1-D5)[1], the function of performance measurement in the form of exams, tests, homework, and so on includes the following:

- A guidance function is used to determine the learner's current level and take into account which explanations and learning content are required.
- A diagnostic function is used to recognize progress and identify weaknesses.
- A performance assessment is used to determine what the learners have actually learned, to assess or recognize the learner's level of knowledge and performance, and to determine their progress.
- The grading of learners.

Just fewer than 70% of respondents2 see the function of performance measurement as obtaining a grade at the end, while over 50% see performance measurement as a way of obtaining information about their own performance level. Other functions mentioned include obtaining a language certificate (17%) and recognizing one's own learning progress (24%).

The interviewees in the private language courses gave similar answers to those at the university. No fundamental differences can be identified. The teachers in the private language courses consider the functions very important for recognizing the weaknesses and strengths of the examinees and learners, for support (where learners need support), for assessing the level of performance, and for recognizing progress.

C. Quality Criteria

Performance measurement is subject to certain criteria known as quality criteria, namely, objectivity, reliability, and validity.

(a). Objectivity

"This criterion means that the same linguistic performance should be assessed in the same way by all correctors" (Albers et al., 1995, p. 26). Albers et al. (1995) also posit that subjectivity in tests can be reduced if the points for the tasks are awarded on the basis of a standardized scoring guideline. Tinnenfeld (2002; cited in Tinnenfeld, 2013) mentions the difficulties in fulfilling this criterion and emphasizes, among other things, that it is possible to achieve objective tests by precisely describing the relevant test situations in advance. According to the teachers, the criterion of objectivity cannot be fulfilled 100%, but subjectivity can be reduced. For them, objective assessment is difficult, but the defined criteria prevent incorrect assessment. The measures mentioned by the teachers include defining the criteria in advance, carrying out various training sessions on correct assessment, involving a second examiner for assessment, multiple assessments, or splitting tasks for assessment. A similar result was also seen from the learners. When asked whether they were assessed under the same conditions as their fellow peers and whether the assessment was objective, the surveys yielded the following responses, as shown in Figure 1.

[1] The interviewees are divided into three groups and quoted according to these letters: D is for lecturers and teachers at the universities in Kosovo; L is for the teachers at the private language courses in Kosovo; and S is for teachers at a private school where German as a foreign language is taught up to level C1. The interviewees are anonymous and are named after the letter (university, language course, or school) and after a number.

[2] Students at universities, schools, or private language courses who have learned B2 or C1.

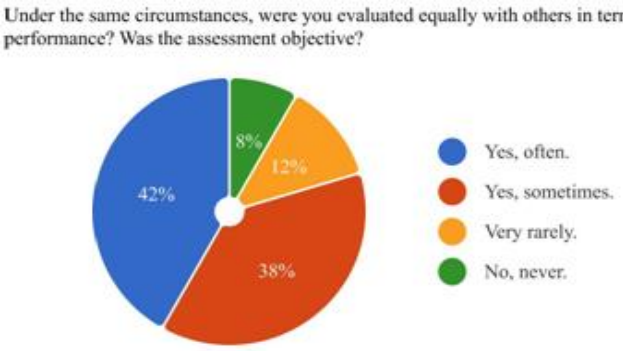

Figure 1. Objectivity in Performance Measurement

Figure 1 shows that 42% of the respondents stated that they were assessed equally in the same circumstances, while 38% of respondents stated that they were sometimes[3] not assessed equally. Moreover, 20% of respondents stated that they were "rarely" (12%) or "never" (8%) objectively evaluated. This indicates that objectivity is largely fulfilled, but not to the expected extent.

(b). Reliability

"The second requirement ... concerns the reliability of the performance measurement, i.e., the reliability of the test. Thus ... a certain linguistic performance will always yield the same score" (Albers et al., 1995, p. 25). According to Tinnenfeld (2013), it is important to note with this criterion that all test items must be offered the same conditions in order to achieve a reliable test result. The interviews showed that it is not possible for learners to demonstrate their performance under ideal conditions. However, it is important that all examinees feel treated equally in the examination room and that no one is disadvantaged. The respondents stated that the criterion of reliability cannot be fully met. Although the option "rarely" (50%) appears, it can be understood that this criterion cannot be fulfilled 100% in reality.

(c). Validity

"The criterion of validity means that a test actually tests what it is supposed to test" (Albers et al., 1995, p. 22). Tinnenfeld (2013) points out that the tasks should not test anything that has not been learned or clearly explained beforehand. The interviews revealed that it "sometimes" (47%) or "very rarely" (40%) happens that the respondents are asked questions that are not covered in class.

D. Forms of Performance Measurement

Glaboniat (2017) provides a definition of the typology of performance measurement and various forms of assessment and explains that diagnostic procedures measure an individual's current language proficiency and establish a differentiated competency profile. Grotjahn and Kleppin (2017) mention several forms of performance measurement. For the purposes of this study, two forms of performance measurement are considered relevant: formal and informal, along with both formative (continuous) and summative (final) assessment.

"Formal forms of evaluation are mostly examinations and tests. Formal examinations and tests are usually the result of lengthy and time-consuming efforts by specialists" (Grotjahn & Kleppin, 2017, p. 33). Examples of such forms of performance measurement include the certification exams offered by the Goethe Institute, ÖSD, and TestDaF, among others. These examinations must meet certain quality criteria, with standardization being a central characteristic. The aim is to make the conditions as comparable as possible in order to ensure valid and comparable results. In contrast to the formal forms of performance measurement, the informal forms are less demanding and require less elaborate processes. However, despite their informal nature, the quality requirements are still inherent. A typical example of informal performance measurement is proficiency testing based on learning objectives.

"The assessment of performances can occur either punctually and product/result-oriented at the end of a learning phase (summative) or continuously and process-oriented integrated into the teaching (formative)" (Grotjahn & Kleppin, 2017, p. 35). According to the students and learners surveyed, continuous (formative) assessment in the form of homework, various projects, or seminar papers takes place "often" (25%) or "sometimes" (26%). The rest of the respondents indicate that this form of assessment is either "rarely" (39%) or "never" (10%) used. When asked whether they prefer continuous or final assessment, 78% of the respondents stated a preference for continuous assessment. Conversely, 22% prefer to be assessed only in the final exam.

Summative (final) assessment is generally used at universities, as indicated by both the questionnaire and the interviewees, although formative assessment is preferred and used whenever possible in the form of projects, seminars, or homework. Summative assessment is a requirement of the university and must therefore be offered to students. It is also used by examination centers such as the Goethe Institute, ÖSD, or TELC.

According to the respondents, the most commonly used forms of assessment (over 75%) are tests, exams, or colloquia,

[3] Respondents had these options to choose from: *often, sometimes, very rarely,* and *never.*

which are essentially a form of summative assessment. Other forms of assessment are used only "sometimes" (projects, seminars) or "rarely" (homework) for evaluation.

E. Types of Tasks

There are different types of tasks within the various forms of performance measurement, and the difficulty level of performance measurement varies for each of these types of tasks. According to Albers et al. (1995), these are classified as open, semi-open, and closed tasks. A task is considered open if the answer can be given relatively freely and must be formulated by the candidate as a productive performance, whereas in closed tasks, learners must select the correct answer from the provided options (Albers et al., 1995).

The three interview groups indicated that it is easier to correct closed tasks. While open tasks are more difficult to assess, problems with evaluation did not arise or seldom occurred. When asked whether the learners "often" had the feeling that they were not assessed correctly in open or closed tasks, it was found that the majority of them were "sometimes" or "rarely" not assessed correctly in open tasks. In contrast, this either "never" or "rarely" occurred with closed tasks (see Table 1).

TABLE 1
ASSESSMENT OF EVALUATION DISPARITY BETWEEN OPEN AND CLOSED TASKS BY LEARNERS

Respondents	Often	Sometimes	Rarely	Never
Open tasks	24%	41%	27%	8%
Closed tasks	4%	21%	30%	45%

IV. MEASURING PERFORMANCE IN THE SKILL OF WRITING

Previously, the focus of this research has been mostly about defining performance measurement in general, explaining the criteria, explaining the forms of performance measurement, and explaining the types of tasks. Now, the focus shifts to the role of writing skills and error correction in the assessment of this skill.

The tasks in foreign language examinations vary in structure and adhere to criteria largely determined by the Common European Framework of Reference for Languages (CEFR), which "provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc., across Europe." (Council of Europe, 2002, p. 14).

A. The Role of Writing Skills

In the development of textbooks in the 1980s, the skill of writing played a subordinate role compared to other skills such as speaking, reading, or listening. It was only later that the skill of writing became more important in the classroom (Bolton, 1996). Storch (1999) emphasizes the importance of writing in German as a Foreign Language (GFL) instruction within the context of pedagogical orientation. He points out that different types of learners have varying preferences for oral or written expression. Particularly for learners interested in writing, writing serves as a suitable exercise for psychological reasons. Therefore, "teaching should provide these learners with sufficient opportunities to express themselves in writing according to their preferred learning style" (p. 249).

According to Rösler (2012), writing serves as a significant tool for identifying competence deficits. In today's communication-oriented society, writing is not only a skill but also a crucial competence, as Kic-Drgas (2022) describes. Communication skills play a prominent role, especially in professional contexts. The use of specialized terminology, the precision of expression, and the selection of appropriate language registers are just a few of the many elements that are taken into account in professional written communication. Therefore, it is evident that writing skills should be specifically promoted in the classroom. Kast (1999) describes writing as a versatile, media-related skill. It serves communicative, instructional, and learning-psychological needs, supports the structuring of thoughts, and enables the conscious development of thought when writing. Two central aspects are the concretization of thoughts and the slowing down of processes, which allow room for reflection and structuring.

Essentially, the skill of writing in the classroom serves two different functions. Firstly, writing serves as a means to an end, e.g., in written exercises to consolidate vocabulary or grammatical structures or when completing written homework. Secondly, writing plays an important role as an independent skill, as it is consciously trained and practiced.

The reasons for neglecting writing skills include a lack of time in the classroom, the hurriedness of curriculum planning, a focus on communication that is predominantly on speaking, and the great amount of effort required for correction, among others. According to the respondents, writing is frequently practiced, with more than 65% stating that they do so "regularly" and 24% saying they do it "sometimes". However, it is important to point out that the skill of writing is not necessarily pursued as the primary objective, but rather that writing serves as a means to an end. When asked about the consideration of writing as an independent skill in the classroom, the results revealed that writing is only "sometimes" used as an independent learning objective in 59% of cases or "rarely" (22%). Teachers from both schools and private language courses confirmed that writing is practiced and trained in both cases. However, the results of the interview and the questionnaire suggest that more emphasis should be placed on the skill of writing.

B. Error Correction and the Evaluation of a Writing Task

A writing task contains specific criteria on the basis of which the candidate's performance is assessed. For the purposes of this work, criteria from various authors, different examination centers such as Goethe, ÖSD, and TELC, and the interviewees (examiners) are used as the basis. Before delving into evaluation and correction, it is crucial to define the term “error”. Kleppin defines errors as "something that violates or deviates from something perceived as correct" (1998, p. 14).

Errors are classified according to specific levels of language. According to Kleppin (1998), from this perspective, errors can be categorized as follows:

- Phonetic or phonological errors (pronunciation or orthography).
- Morphosyntactic errors (morphology and syntax, e.g., verb conjugation, sentence structure, etc.).
- Lexico-semantic errors (incorrect word usage in context or changes in meaning).
- Pragmatic errors (stylistic inconsistency, culturally inappropriate behavior).
- Content errors (statements expressed incorrectly in terms of content, e.g., "Berlin is located in southern Germany") (p. 43).

Another distinction between error types is based on the degree of difficulty, in the sense of how severe an error is. Therefore, errors are differentiated as either major[4] or minor[5] errors (Kleppin, 1998). Moreover, the correction of errors in written assignments can vary. According to Kleppin (1998), there are four types of corrections in written assignments:

1. The examiner or teacher underlines the incorrect part without going into detail about what is incorrect.
2. The examiner or teacher underlines the error and marks the error; for example, whether it is a lexical error or an error in genus, tense, etc.
3. The examiner or teacher marks the error and writes the corrected version of the utterance.
4. The candidate or learner either corrects the error in the error marker or rewrites the task set by the examiner/teacher.

From the three interview groups, it emerged that all types of correction are used. This depends on the specific institutional guidelines, the examiners themselves, or other factors (e.g., time constraints, group size, and so on). The figure below shows the most commonly used form of correction based on learners' experiences.

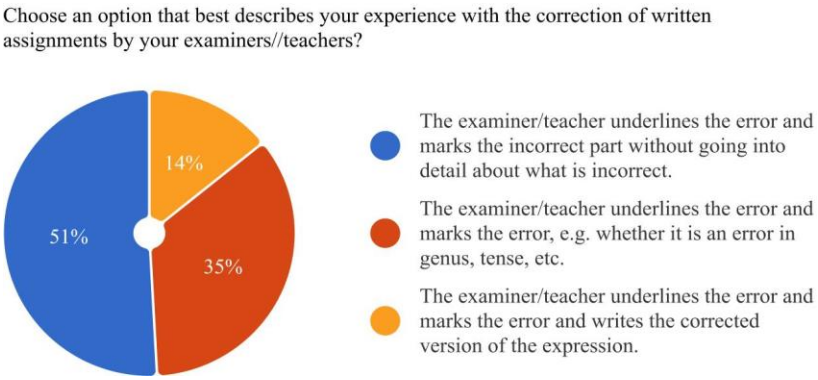

Figure 2. Forms of Correction by Examiners/Teachers

The majority of respondents (51%) stated that errors were only underlined without going into detail about what was incorrect. The second form of correction also occurred, with 35% of respondents receiving such corrections from their examiners or teachers. The least used form of correction was the third, as reported by 14% of respondents.

Despite the common goal of evaluating a topic, there were considerable differences in the evaluation criteria used by the respondents. For this work, the evaluation criteria provided by Tinnenfeld (2013) were used as written below:

- Linguistic correctness (communication, audibility, compliance with grammatical norms, assessment of the language behavior of native speakers, suitability for study and work, reference to individual language structures).
- Orthography.
- Expressiveness (use of appropriate technical terms, text-linking elements, correct use of written relevant constructions, consideration of hypotactic constructions).
- Content organization (structural coherence of the text, strength and relevance of the cited arguments and examples, presence and nature of logical connections, authenticity of the text genre, consideration of inter-thematic relations).

While the evaluation criteria in the language centers (e.g., Goethe, ÖSD, and Telc) are evenly distributed according to points or percentages, the criteria mentioned by the interviewees differ in their distribution of points. Upon comparison, it is evident that although the nomenclature may vary, they assess the same aspects.

[4] A violation of one or more of the language levels mentioned above can lead to misunderstandings among native speakers.
[5] Linguistic phenomena that have not yet been practiced or are barely audible when reading the sentence aloud.

The criteria outlined by Tinnenfeld (2013) are also grouped similarly by the interviewees for easier presentation and analysis, as the interviewees also apply four criteria for evaluation:

- Content organization (task fulfillment, content completeness, content, overall impression) was mostly rated with 3 out of 10 points (30%).
- Expressiveness (expression, vocabulary) was mostly rated at 2 out of 10 points (20%).
- Linguistic correctness (text coherence, text structure) was rated at 2 out of 10 points (20%).
- Orthography (grammar, grammatical, and linguistic correctness) was mostly rated at 3 out of 10 points (30%).

As can be seen from the interview data, the focus of the evaluation is on content (content structure) and grammar (orthography). However, a detailed analysis shows that content plays a more important role for the interviewees than grammar. One statement in particular from interviewee D2 exemplifies this view:

> "The content is important with regard to whether the task is fulfilled. The fulfillment of the task is the first aspect we consider. If this criterion is not met, the text will not be corrected and will not be considered for correction. Task fulfillment comes first, but once this criterion is met and the word count is also met, we move on to evaluating the individual criteria" (D2).

A similar statement was also revealed by another interviewee (D3) when asked which of the two aspects, "grammar" or "content", is more important:

> "You cannot have one without the other. You cannot have good content without good form. For me, this is a fundamental truth of language—the separation of content and grammar that we as language teachers have concocted. But in reality, there is only one language, and it has a certain content. And when I correct, I have to say at the same time: What you are saying is wrong. When I correct, I look at the grammar and the structure, but I mainly try to intervene where I can show that the writing is actually pursuing a certain content that it cannot grasp because of this incorrect grammar, i.e., I always try, if possible, to show that the semantics are to a certain extent a result of the syntax. When I correct a text, I read it through and simply try to understand what it is about and, above all, the content. I look at the overall impression of how much courage the person has to convey sophisticated content or whether they are writing trivial things. That's the first thing I do. And then, on the second or third pass, I look at the individual errors, but I believe that the first impression is more important" (D3).

It can be concluded from the interview data that content takes priority in the writing evaluation process. Only after successfully completing the task and meeting the word count do other criteria come into focus.

C. Sources of Error in Evaluation and Countermeasures

The difficulties of performance measurement concern not only the determination of the criteria but also other factors that are not directly related to the performance of the examinees. Examiners are consciously or unconsciously influenced by subjective factors. Subjectivity in assessment is recognizable; for example, it can be based on first impressions, based on the overall assessment of a task, or the opposite, and so on. From the analysis of the interviews with interviewees from universities, private language courses, and schools, it can be seen that errors in assessment cannot be ruled out. Countermeasures mentioned by interviewees include checking the tasks again, assessing the tasks with a second examiner, setting the criteria in advance, and so on. Another important factor is experience with the assessment. Initially, the tendency to make mistakes in the assessment is higher than with many years of assessment experience.

The errors in the assessment were also noted by the interviewees based on their experiences as learners and students. The respondents' impressions of the assessment of speaking and writing skills can be compared here. The responses to the questionnaire indicated that respondents often felt that they were not assessed correctly in an oral examination due to subjective factors. Figure 3 shows that more than half of the respondents (20% with one[6] and 23% with two) often felt that they were not assessed correctly in an oral examination.

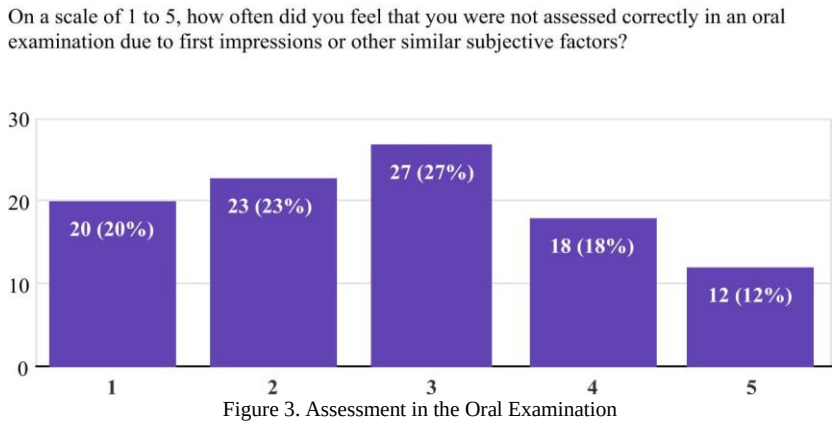

Figure 3. Assessment in the Oral Examination

[6] Respondents had to choose from a scale of 1 to 5: Very Often 1 - 2 - 3 - 4 - 5 Never.

This feeling is not as strong when assessing a written exam. Figure 3 shows that, in comparison with the oral examination, only 19% (9% with one and 10% with two) felt that they were not assessed correctly in a written examination.

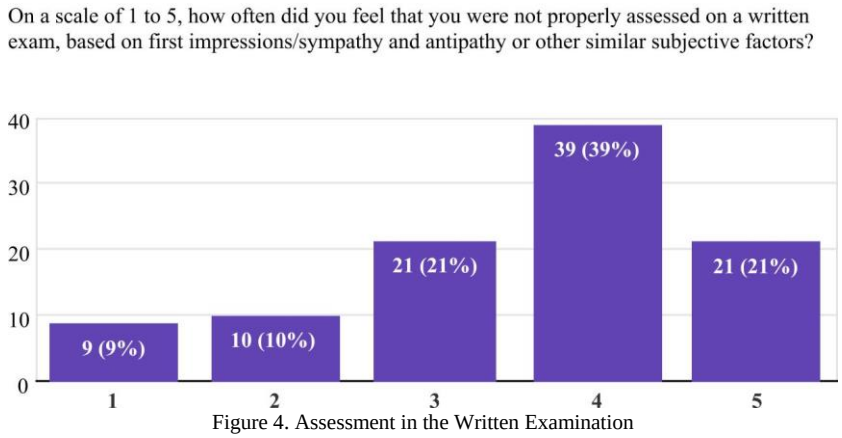

Figure 4. Assessment in the Written Examination

According to the interviewee, writing tends to be assessed more strictly than speaking. When assessing the skill of writing, much more attention is paid to errors, and the assessment is more precise as the text can be read and corrected multiple times. Conversely, in speaking, both the evaluation and examination occur simultaneously, as emphasized by interviewee D3:

> "It's a very difficult question. I believe I would tend to say that the likelihood of making an incorrect assessment is greater in writing because it often becomes meticulous and does not focus on the whole: what does the text express and what can it convey? On the other hand, when speaking, you have someone in front of you, and you don't pay as much attention to each individual word, so the overall message comes across much better. This doesn't happen in writing because one tends to focus too much on the individual error in the word or sentence. At first glance, I would say the likelihood of me being too strict in my evaluation or making the person look bad is greater in writing" (D3).

From this perspective, it is evident that errors in written tasks are assessed more strictly than errors in speaking. However, this should be considered in general and not specifically with regard to objectivity. When asked about the differences in evaluations between speaking and writing, it can be assumed that subjective factors are more visible to examinees in speaking evaluations than in writing. In addition, the criteria for written tasks are provided during the examination, as confirmed by the majority of respondents, as shown in Figure 5. Over 58% of respondents indicate that the criteria for tasks are provided, while 10% believe that examiners do not adhere to these criteria. Overall, it can be observed that the evaluation of writing tasks is considered more objective and accurate compared to tasks in an oral examination.

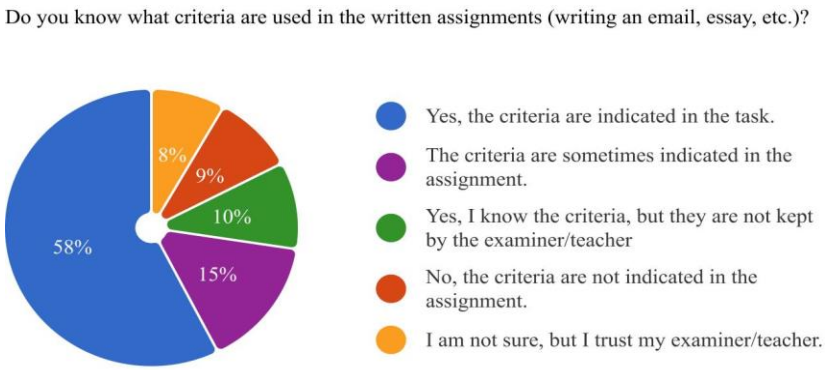

Figure 5. Criteria Entry for the Written Task

The interviews also revealed that, while writing is corrected extensively as it can be read and reviewed multiple times, this is not the case with speaking. Errors are more difficult to detect as assessment and speech production run parallel to one another.

Despite the large number of specifications and criteria, performance assessment is a challenge. It is clear what difficulties both examiners and examinees are confronted with during assessment; compliance with quality criteria is held in high regard, particularly. Of great significance and importance for this study were the various forms of correction, the assessment criteria, as well as the typology of tasks and the role and functions of writing skills. The detailed discussion

of all these aspects on the basis of the literature, interviews, and survey has enabled a more precise presentation of the topic by linking them together in a meaningful way.

V. Conclusion

The aim of this article was to find out how performance is assessed, what difficulties are expected in this process, and what criteria this assessment is based on. It has been shown that certain quality criteria must be met for a justified assessment, including objectivity, reliability, and validity. In reality, however, these criteria are not fully met. From the respondents' perspective, objectivity is only "often" fulfilled in 42% of cases, while 38% of respondents stated that it is "sometimes" fulfilled, a further 12% said "very rarely", and 8% even went so far as to say it is "never" fulfilled. While the percentage for not meeting the objectivity criterion was high to a certain extent, the two other criteria were significantly better. As half of the respondents indicated, unequal treatment of conditions for all examinees (e.g., health impairments, lack of concentration, ensuring equal conditions for all examinees) in examination situations occurred only "rarely". With regard to the third criterion (validity), it was found that 40% of the respondents were "sometimes" asked questions in the exam that were not covered in class.

The forms of performance measurement were also a central theme of this study. In many forms of performance measurement, institutional guidelines play a role. Performance measurement was followed either formally or informally, with the aim of assessment usually being a formal evaluation. While summative assessment is predominantly used by institutions and examination centers, formative assessment is desired and recommended both from the perspective of the respondents and from the perspective of the interviewees. While respondents either "never" (45%) or "rarely" (30%) felt correctly assessed for closed tasks, the situation is quite different for open tasks. In that case, 41% of respondents indicated that they "sometimes" felt incorrectly assessed in the open tasks. Moreover, the role of writing skills was also researched in this article. It was found that a significant amount of writing takes place in the classroom. 65% of respondents indicated that they "often" write in class, in the sense of doing exercises or taking notes. However, when it comes to practicing writing as a target skill by writing letters, emails, or essays, respondents indicated that they "sometimes" write (59%). This fact was also confirmed by the interviewees. Textbook planning played a role, as did time constraints. Writing tasks take up much more time, both to create the texts and to correct them. Speaking proficiency is significantly more important than writing proficiency. Over 75% of respondents shared this opinion.

Regarding assessment, conscious or unconscious biases on the part of examiners became apparent. After numerous steps and instructions, it is crucial to review the assessment several times or to have the tasks corrected by two or more examiners. A notable result of the questionnaire was that objectivity was better achieved in written tasks than in oral tasks. Furthermore, the interviews revealed that written tasks can be assessed more rigorously than oral tasks.

To meet the criteria for assessment, specific criteria are provided for levels B2 and C1. Both in the interviews and in the model tests from institutions such as Goethe, ÖSD, and TELC, there were typically four main criteria. These criteria were divided into four categories: content organization, expressiveness, language accuracy, and grammar (orthography). While the assessment criteria are equally present in the tests from language centers (in most of the model tests), the distribution of the criteria varies between the respondents. Content organization and grammar were assigned more points than expressiveness and language accuracy. However, according to respondents, the criteria for assessing writing tasks are not always specified. Only 58% of respondents indicated that criteria were provided for the tasks. As analyzed throughout the study, it is evident that the assessment criteria used may vary in their naming but ultimately assess similar or even identical aspects. This finding allowed for the grouping of these criteria into four overarching categories to achieve a clear and comprehensive result.

The compiled interview results make it clear that content organization and orthography are perceived as particularly significant criteria for the assessment of language skills. This underlines the consistency and standardization of the assessment approaches used by the interviewees. The other criteria only come into focus once the task has been successfully completed and the specified word count has been met. The content is considered more important. However, when the total number of points is divided by the number of points in four criteria areas, it is evident that content and grammar, or orthography, are considered to be of equal value. The hypothesis can therefore be partially confirmed. Although content is fundamental to the writing task, as mentioned by one of the interviewees, "There is not one without the other".

In conclusion, the picture that emerges is one of a challenging performance assessment from the perspective of both the examinees and the examiners. Despite the existing guidelines and quality criteria for the assessment of open tasks, there are deviations in assessment practice. Although there are proven measures to minimize assessment errors, their implementation in practice is sometimes difficult due to a variety of influencing factors. Reality does not always allow for ideal implementation.

References

[1] Albers, H., Bolton, S., & Jenkins- Krumm, E. (1995). *Testen und Prüfen in der Grundschule* [Testing and Assessment in Primary School]. Fernstudieneinheit 7. Langenscheidt.

[2] Bolton, S. (1996). *Probleme der Leistungsmessung* [Problems of performance measurement]. Langenscheidt.

[3] Decke-Cornill, H. & Küster, L. (2015). *Fremdsprachendidaktik. Eine Einführung* [Foreign Language Didactics. An introduction]. Narr Francke Attempto.
[4] Dudenredaktion (Eds.). (2015). *Deutsches Universalwörterbuch* [German Universal Dictionary]. Duden.
[5] Europarat. (2002). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen* [Common European Framework of Reference for Languages: learning, teaching, assessing]. Langenscheidt.
[6] Glaboniat, M. (2017). *Testen, Prüfen und Beurteilen von Deutschkenntnissen und der Gemeinsame Europäische Referenzrahmen GER* [Testing, Checking and Assessing German language skills and the Common European Framework of Reference for Languages CEFR]. In W. Ulrich, (Ed.) *Deutschunterricht in Theorie und Praxis* [German lessons in theory and practice]. Schneider Verlag, Hohengehren.
[7] Grotjahn, R. & Kleppin, K. (2017). *Prüfen, Testen, Evaluieren* [Checking, Testing, Evaluating]. Ernst Klett Sprachen GmbH.
[8] Hinger, B. & Stadler, W. (Eds.) (2018). *Testen und Bewerten fremdsprachlicher Kompetenzen. Eine Einführung* [Testing and Evaluating foreign language competences. An introduction]. Narr Francke Verlag (Narr Studienbücher).
[9] Kast, B. (2003). *Fertigkeit Schreiben* [Writing skill]. Langenscheidt.
[10] Kleppin, K. (1998). *Fehler und Fehlerkorrektur* [Errors and error correction]. Langenscheidt.
[11] Kic-Drgas, J. (2022). *Entwicklung der Schreibkompetenz in einer Fremdsprache an der Hochschule. Konzept für die Schreibvermittlung im berufsbezogenen Unterricht am Beispiel von Deutsch als Fremdsprache* [Developing writing skills in a foreign language at university. Concept for teaching writing in work-related lessons using the example of German as a foreign language]. V&R unipress.
[12] Langenscheidtredaktion (Eds.). (2008). *Langenscheidt Großwörterbuch Deutsch als Fremdsprache: das einsprachige Wörterbuch für alle, die Deutsch lernen* [Langenscheidt Grand Dictionary of German as a foreign language: the monolingual dictionary for anyone learning German]. Langenscheidt.
[13] *Modelltest Goethe B2*. Retrieved on November 16, 2022: https://www.goethe.de/pro/relaunch/prf/materialien/B2/b2_modellsatz_erwachsene.pdf
[14] *Modelltest Telc C1*. Retrieved on November 10, 2022: https://www.telc.net/sprachpruefungen/zertifikatspruefung/deutsch/telc-deutsch-c1/
[15] Ponsredaktion (Eds.). (2008). *PONS Großwörterbuch Deutsch als Fremdsprache: Ca. 77000 Stichwörter und Wendungen* [PONS Grand Dictionary of German as a foreign language: Approx. 77000 headwords and phrases]. Klett.
[16] Rösler, D. (2012). *Deutsch als Fremdsprache. Eine Einführung* [German as a foreign language. An introduction]. Springer.
[17] Storch, G. (1999). *Deutsch als Fremdsprache. Eine Didaktik. Theoretische Grundlagen und Unterrichtsgestaltung* [German as a foreign language. A didactic approach. Theoretical foundations and lesson design]. UTB.
[18] Suli, L. (2022). A Research on the Blended Evaluation Mode in College English Writing Course. *Journal of Language Teaching and Research*, *13*(4), 763-771. DOI: https://doi.org/10.17507/jltr.1304.09.
[19] Tinnenfeld, T. (2013). *Dimensionen der Prüfungsdidaktik. Analysen und Reflexionen zur Leistungsbewertung in den modernen Fremdsprachen.* [Dimensions of examination didactics. Analyses and reflections on performance assessment in modern foreign languages]. COD.

Sadije Rexhepi earned her PhD in linguistics at the University of Pristina, Kosovo (part of research at Humboldt University-Berlin, on a DAAD Scholarship). She works as an Associate Professor for German Language at the Department of German Language and Literature at the Faculty of Philology of the University of Pristina and teaches German grammar, text linguistics, text analysis, academic writing, testing, and evaluation. Professor Rexhepi has a number of publications in journals on the topics of contrastive linguistics, text linguistics, German linguistics, etc. (ORCID ID: https://orcid.org/0000-0003-0478-7560)

Çliresa Suka earned her Master in linguistics at the Department of German Language and Literature at the Faculty of Philology of the University of Pristina, Kosovo. In 2022/2023 she had an internship with Erasmus+ program at Humboldt-University of Berlin, and currently she is in an internship at the German Parliament in Berlin. MA Suka is a PhD candidate at the Humboldt University Berlin, Faculty of Linguistics and Literature. (ORCID ID: https://orcid.org/0009-0001-6098-3200)