Znanstveni pregledni članak
UDK 371.33:784.4:811.131.1'243

Primljen 13. 1. 2008.
Prihvaćen 23. 4. 2008.

ASSESSING WRITING THROUGH BENCHMARKS: A CASE STUDY IN THE SLOVENIAN CONTEXT

*Mihaela Zavašnik, Karmen Pižorn**

Sveučilište u Ljubljani

Assessing writing scripts in high stakes examinations has long attracted test designers, test evaluators and other researchers in testing languages. There are many various and interdependent elements that may affect the scoring procedure of writing scripts. Rating scales which include levelled descriptors which further consist of specific testing terms are of paramount importance. If the wording of the rating scale is not perfectly comprehensible and equally applied, it will lead to unreliable results and unfair scoring of writing tasks. To decrease subjectivity of scoring writing scripts a procedure of constructing and implementing benchmarks for each band of the valid rating scale should be established.

Key words: language assessment, language testing, assessing writing, benchmarking, rating scale, high stake examination

1. INTRODUCTION

Valid and reliable assessment does not necessarily imply automatic comprehension of performance or content standards that test takers should achieve in proficiency and achievement (high stake) examinations. This is especially true for written assessment which should be a part of any modern language test, and especially for marking schemes incorporating rating criteria, rating scales, descriptors, glossaries, etc. If the terms in rating scales are not equally applied, adequately understood, and supported by anchor writing scripts, then any valid and reliable interpretation of the test results is obstructed to such an extent that all the resources, financial and academic, have been wasted.

In this article we will introduce the benchmarking procedure as it has evolved through the last decades and across diverse disciplines, and especially within language testing community. A detailed step-by-step description of the benchmarking process in written assessment in the context of Primary School-Leaving Examination in English Language (PSLEE) is followed by anticipating solutions to the questions why written justifications for scripts which exemplify bands for each criterion of the rating scale should be developed.

* Mihaela Zavašnik, Sveučilište u Ljubljani, Filozofski fakultet, Ljubljana, e-mail: mihaela.zavasnik@guest.arnes.si

2. SECOND/FOREIGN LANGUAGE WRITING

Writing is essential to thinking and learning. As a strategic process, writing is a key element of communication, a critical part of comprehension and an important problem-solving activity. Writers use a range of skills and strategies in the process of writing as they communicate with diverse audiences and for many different purposes. It has long been recognized that writing is a complex activity and that it is more than just a skill or talent. As such it is a means of exploration and manifestation for learning in all grades and disciplines of educational processes. There are still many issues that the educationalists keep pondering on, for example, how the process of the writer in the real world can be developed in the classroom, how writing can be taught across the curriculum, and how writing can be fairly and authentically assessed.

Cushing Weigle (2002: 35) claims that

> *'over the past years a consensus has emerged among researchers that second-language writers use many of the same writing processes in their second language as in their first, and expertise in writing can transfer from the first to the second language, given at least a certain level of language proficiency'.*

However, writing in second/foreign language may be hampered because of the need to focus on language rather than content. Silva (1993: 668) found out that writing in a second language tends to be *'more constrained, more difficult and less effective'* than writing in a first language. He (ibid.) also claims that second language writers *'plan less, revise for content less, and write less fluently and accurately'* than first-language writers. We can see that differences between first and second/foreign-language writing are considerable, and in particular the variety of backgrounds, experience, needs, and purposes for writing is much greater for second/foreign-language writers than for first language writers. This variety has important implications for the assessment of second/foreign language writing, both in terms of designing appropriate writing tasks and in terms of evaluating writing.

3. ASSESSING AND SCORING SECOND/FOREIGN LANGUAGE WRITING

Writing as a communicative skill may be assessed in different ways supported by various and sometimes diametrically opposed theoretical approaches to teaching and assessing writing. In American testing context the so-called *indirect tests* of writing were commonly used a few decades ago. These most often referred to multiple choice tests of grammar and usage. Another view, proposing a collection of written texts written for different purposes over a period of time, is represented by Williamson (1997: 237–238) when he claims that

'Writing is an extremely complex phenomenon that can only be understood through such 'messy' assessment techniques as **holistic scoring and portfolio assessment**. . . . *The richer assessment techniques, although messy to collect and evaluate, make much more sense than the more orderly multiple-choice tests of writing ability precisely because of the richness of the information they reveal to both researchers and teachers. Unfortunately our arguments continue to fly in the face of the logic of such theoretical categories from the psychometric literature as reliability and validity, as they are conventionally, and thus, narrowly defined.'*

The third approach to writing assessment is *timed impromptu writing test* (Cushing Weigle, 2002) which is a kind of more or less satisfactory alternative to indirect tests of writing, on the one hand, and portfolio assessment on the other. A timed impromptu writing test is about a required topic in a specified amount of time. Students are given a topic through writing prompt and the task has to be completed in a short amount of time. Research and experience indicate an increase in the efficiency and reliability of writing tests on condition that the prescribed procedures are followed strictly and consistently. However, there has been some criticism about the validity of this method of testing and especially whether the procedures that lead to scoring reliability actually detract from validity (Charney, 1984; Huot, 1990 and 1996 in Cushing Weigle, 2002).

An important, if not the most important part of the assessment of writing is the scoring procedure. The scoring procedures are critical because the score is ultimately what will be used in making decisions and inferences about writers. The score is '*the outcome of an interaction that involves not merely the test taker and the test, but the test taker, the prompt or task, the written text itself, the rater(s) and the rating scale'* (Hamp-Lyons, 1990 and McNamara, 1996 in Cushing Weigle 2002: 108). Of the named elements, two are of central consideration in scoring: defining the rating scale and ensuring that raters use the scale appropriately and consistently. The consistent application of the scoring rubric is considered essential to the validity and meaningful interpretation of scores (Messick, 1994). Bechmarking helps to define the standards of performance for a given assessment and serves as the rubric's surrogate reference points.

4. BENCHMARKING AS S SCORING PROCEDURE FOR WRITING ASSESSMENT

Benchmarking was originally developed by companies operating in an industrial environment. It has therefore been applied most widely at the level of the business enterprise. In recent years, organisations such as government agencies, hospitals and universities have also discovered the value of benchmarking and are applying it to improve their activities and systems. In other words, benchmarking is the well-respected business practice of comparing one organization's procedures with those of another whose practices are exemplary. While this is a relatively new idea in education, the benefits are similar.

4.1 What is a benchmark in language testing?

Research regarding the role of benchmarks in scoring writing assessments is surprisingly limited. There have been some projects carried out on in Europe (Tanko, 2004), however, relatively scarce information is available in writing. In this part of the article, the authors will for the mentioned reason address the issue/topic mainly from their subjective view, experience and knowledge rather than any objective available information.

Benchmarking may be defined as a specific level of performance that is expected in a given subject, in a given grade, and/or in a given writing criterion. A benchmark is usually a set measurement point used to assess whether students are progressing toward a specific goal. Benchmarking should not be considered a one-off exercise. To be effective, it must become an ongoing, integral part of a continuing improvement process with the goal of keeping abreast of ever-improving best practice.

An illustrative example of benchmarking is *Canadian Language Benchmark System* (see http://www.language.ca) which describes a person's ability to use the English language to accomplish a set of tasks. The Canadian Language Benchmark document (ibid.) provides detail about what it means for a learner to be level 4 in writing. Criteria tell the what. The how well is called performance standards. Finer grained than the more superordinate content/performance standards, the standards in marking schemes are specific descriptions of performance levels and requirements of details expected in the performances of test-takers. In the Canadian case the marking scheme is presented as a matrix in which the criteria and the standards form horizontal and vertical axes. For each cell in the matrix, an anchoring description is provided and a student paper is offered as a benchmark to help scorers match performance levels for each criterion with test-takers' writings.

Frequently referred to as scoring guides, rating scales are rules by which the quality of answers is determined. Each rating criterion can be defined as a description of student performance that clearly articulates the requirements for each of the score points. The use of rating scales with their incorporating rating criteria and descriptors predates the current standards movement, but wording of rating scales is crucial in current implementations of academic standards. As such, benchmarking may be described as an unavoidable stage in a rating procedure of any writing test. To achieve the best possible outcome of this procedure, there are a certain number of inevitable activities that have to be performed.

Firstly, the writing task that will be used in the benchmarking process must incorporate a number of features. It should be very carefully designed and allow all test takers to perform to the best of their abilities. It is also supposed to have elicited a range of performances and is a reflection of the latest test specifications. No or few problems should be noticed with the task itself.

Secondly, a representative number of scripts which reflect the various levels given by the scale both in terms of overall grade and for each criterion have to be selected by a test designer, preferably the chief examiner. The chief examiner then draws up written justifications for the selected scripts and grades them according to each rating criterion and calculates the total sum of points for each script. Then s/he delivers these scripts with empty

grids and together with the marking scheme and any guidelines for using the scheme to all the members of a testing team who are asked to grade the scripts and assign individual marks to each criterion.

They complete the grid with their own marks and write justifications for the awarded grades. The grids and written justifications should then be returned to the chief-examiner. Each member is assigned an identification code and all the findings are compiled into one table including the chief-examiner's grades for each criterion, i.e. one table for each script. The chief-examiner identifies the consensus and variation scripts. Then s/he selects ONE consensus script for each level (=band) of each criterion, for example, one script for band 0 for task fulfilment, one script for band 0 for vocabulary, etc.

Practical experience shows that there are certain open issues which the benchmarking process should take into account. These may refer to the following questions:

⇒ How long has the marking scale been used?
⇒ Has any particular problems in using the scale been noticed so far?
⇒ Have all the bands been spread across each of the criteria?
⇒ Have the bands within each criteria been perceived as being a 'pass' performance in that criteria?
⇒ Has any attention been paid to the order in which candidates dealt with the bullets?
⇒ Was there a definitive list of grammatical structures that candidates were expected to 'know' which might help in clarifying such phrases as 'variety of structures'?
⇒ How will the final written justifications with their respective scripts be used in the future?

5. THE ASSESSMENT OF WRITING AND THE APPLICATION OF BENCHMARKING IN THE SLOVENIAN PRIMARY SCHOOL-LEAVING EXAMINATION IN ENGLISH LANGUAGE (PSLEE): A CASE STUDY

PSLEE consists of two parts: written and oral and is based on the objectives defined by The National Syllabus for English in Nine-year Primary school. The written part consists of one paper, which is divided in three sections: Reading Comprehension, Use of Language and Writing.

The aim of the *Writing* tasks (the case studied and described in the paper) is to examine learners' ability to perform written tasks which have a communicative purpose, such as to inform, express one's opinion, narrate, advise, suggest, etc. The PSLEE has adopted the timed impromptu writing test approach to assessing writing as a half of the student's final score in English already consists of classroom-based evaluations of language skills. A student's writing skills are assessed continuously throughout the school years of English instruction and some of the teachers have already implemented portfolios as instruments of writing assessment. However, there is still a lack of empirical research in ELT community which would back up the sometimes self-contained convictions of its proponents about the

role and limitations of portfolios.

Students taking the PSLEE write a guided piece of writing and may be asked to write a postcard, a greetings card, an invitation, a personal letter, a description of a person, an event, etc., a short story, a diary entry or an article (see Appendix A). The writing part consists of one writing task containing 60 to 80 words. Each writing task type provides a stimulus (textual or visual), an addressee, the purpose of the text, the word length of the text and valid assessment criteria. Rubrics are given in the learners' mother tongue, except for the input information that is necessary to complete the task. The use of dictionaries is not allowed.

Assessing of the scripts is performed by centrally trained external raters. As all the tasks are of a subjective type they are assessed to the following criteria: task fulfilment, vocabulary, grammar and organisation. Learners' performance is assessed using a four-level (0–3) analytical rating scale for task fulfilment and vocabulary, and a three-level (0–2) analytical rating scale for grammar and organisation (see Appendix B).

5.1. Problem areas: reasons for benchmarking in the context of the PSLEE

Performance-based assessment relies on strategies and applications of knowledge and skills through the performance of tasks that are purposeful, meaningful and captivating for students. This type of assessment should provide teachers, students and test makers with information about how a child understands and applies knowledge. The benefit of performance-based assessments is well documented. However, many teachers are hesitant to implement performance-based tasks in their lessons and even less in their own designed test papers. Commonly, this is because these teachers feel they do not know enough about how to fairly assess a student's performance.

It is thus not surprising that the Testing Team members for English had felt very reluctant to designing a workable rating scale, verbal scale descriptors and other guidelines for external raters. The most outstanding issue was that the results of inter-rater reliability between the external raters and the chief-examiner, as well as other members of the Testing Team for English had not been very promising after the first live writing task was administered to approximately 150 primary school students across Slovenia. What was even more worrying was the fact that the inter-rater reliability among the Testing Team members themselves had been much too low. Only two out of 47 scripts, which had been graded by 9 members of the testing team, showed a total consensus in all four rating criteria (task fulfilment, vocabulary, grammar and organisation) and on approximately 20% of graded scripts an agreement in a particular rating criterion had been reached. For example 14 (30%) scripts were awarded the same amount of points for task fulfilment. The members were aware of the golden rule that if a test does not offer consistent results then the decisions that are made on the basis of such results will be misleading, inappropriate and unfair to the test takers. At the same time, as reliability is a prerequisite for test validity, the test may be disputed on the fact that it does not measure what it claims to measure.

As can be perceived from the last paragraph the rating procedure had been faulty in some of its stages. To overcome these shortcomings The Testing Team members introduced

a benchmarking process to rating writing tasks.

5.2. The outcomes of the benchmarking procedure

The outcomes of the benchmarking process of the selected scripts have been manifold and have answered many queries and at the same time opened new areas for investigation and improvement.

First of all, many new documents have been designed and the old ones revised and changed appropriately. They include 14 written justifications for sample scripts (see Appendix C), a revised rating scale for writing tasks (see Appendix B), a glossary of the terms used in the revised rating scale (see Appendix D), a detailed report on the issues that will have to be dealt with in the future and a set of recommendations for test designers to consider and adopt.

Who will profit from this procedure? First of all, it is the external raters who will be able to grade the writing scripts more consistently and therefore more reliably. The glossary which incorporates detailed explanations of such terms as adequately, rich, almost, a wide variety, no variety, etc. (see Appendix D) will also enable test raters to grade the writing scripts faster. Moreover, these documents will guide the test makers in designing more relevant tasks that is the tasks which will lend themselves to fair assessment and will encourage a wide range of responses. The students and their class teachers who are entitled to inspect their completed and graded tasks will find in the revised rating scale together with the glossary a more helpful and a more sufficient justification instrument for the rewarded grade.

The written justifications for scripts which exemplify all bands in each criteria will contribute to a more consistent and reliable assessing of the writing scripts. A sample of the written justifications is included in Appendix C.

In addition, the benchmarking procedure has identified 'problematic' rating issues, such as 'length' and envisaged a possible solution. Length has been a problem with many candidates writing more than the word limit. Some of the test designers feel that the length should be extended to accommodate this problem but this idea needs to be considered carefully, so to ensure that quality is not being sacrificed to the needs of quantity. One way of ascertaining candidates' perceptions regarding the appropriate length for a writing task could be by way of a questionnaire.

The benchmarking procedure has outlined another characteristic of rating scales, whose paramount mission is to decrease subjectivity in written assessment. This feature refers to the fact that ambiguous and misleading words incorporated in rating scales lead to inconsistent scoring and at the same time cloud the meaning of the performance criteria. Another recommendation concerns the need for a re-running of benchmarking process which should involve another set of scripts from another task.

6. CONCLUSION

To conclude, benchmarking in written assessment allows test raters and test designers to establish a baseline for student performance, ensures accurate feedback on adequate progress

of students in writing skills, ensures sound data-driven decisions on professional development of test making process, as well as classroom instruction, and provides data for improvement in teaching writing skills. Results of the case study in the context of PSLEE demonstrates that the benchmarking process was worthwhile implementing. However, additional studies and research within ELT community are needed to better understand the role of benchmark as a critical element in writing assessment and to evaluate its strengths and weaknesses.

LITERATURA:

- Alderson, J. C., Clapham, C., Wall, D. (1995): *Language Test Construction and Evaluation*. Cambridge, CUP.
- Charney, D. (1984): *The validity of using holistic scoring to evaluate writing,* Research in the Teaching of English, 18, 65–81.
- Canadian Language Benchmark System. Available: http://www.language.ca/ (10 February 2007)
- Common European Framework of Reference for Languages: Learning, Teaching, Assessment. (2001). Cambridge, CUP.
- Cushing Weigle, S. C. (2002): *Assessing Writing.* Cambridge, CUP.
- Huot, B. (1990): *Reliability, validity, and holistic scoring: what we know and what we need to know,* College Composition and Communication, 41, 2, 201–213.
- Huot, B. (1996): *Toward a new theory of writing assessment,* College Composition and Communication, 47, 549–66.
- Messick, S. (1994): *The interplay of evidence and consequences in the validation of performance assessment,* Educational Researcher, 23, 2, 13–23.
- Popp Osborn, S., Ryan, J. (2002): *The effect of benchmark selection on the assessed quality of writing.* ERIC Database: ED481664. Available 10 February 2007.
- Silva, T. (1993): *Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications,* TESOL Quarterly, 27, 657–677.
- Tanko, G. (2004): *Into Europe. Prepare for Modern English Exams: The Writing Handbook.* Budapest, Teleki Laszlo Foundation and British Council.

• Williamson, M. (1997): *Pragmatism, Positivism, and Program Evaluation.* In: Blake, Y.K. and Huot, B. (eds.). Assessing Writing Across the Curriculum: Diverse Approaches and Practices. Greenwich, Conn., Ablex.

Appendix A: THE TASK USED FOR BENCHMARKING

Section III: Writing **Recommended time: 15 min**

Write an answer to the letter which your uncle from Australia has just sent. He intends to visit Slovenia shortly. In your letter write about:

- the places you want to show him
- the clothes he should bring for his visit
- the present you wish to get from him.

Write at least 60 and not more than 80 words. Marks will be awarded for task fulfilment, vocabulary, grammar and organisation.

The letter has been started for you. Continue.

Dear uncle,

Thank you ..□

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....□

(10 points)

Appendix B: REVISED RATING SCALE FOR WRITING TASKS

Task Fulfilment

3 The writing **fully** addresses the writing task
The content is relevant to the task and all the bullets are **developed.**

2 The writing adequately addresses the writing task.
The content is mostly relevant to the task set and almost all the bullets are **developed.**

1 The writing **only partly** addresses the writing task.
The content is **partially relevant** to the task set and/or only one bullet is **developed.**

0 The writing fails to address the writing task or **no answer is attempted.**
The content is irrelevant to the task set **and/or no bullet is developed.**

Vocabulary

3 Vocabulary is rich, adequate and appropriate for the task.
Almost no sentences contain errors of spelling. **These errors do not hinder comprehension.**

2 Vocabulary is adequate and appropriate for the task.
A few sentences contain errors of spelling. **These errors rarely hinder comprehension.**

1 Vocabulary is **partially adequate**, of narrow range and/or repetitive.
Some sentences contain errors of spelling. **Some of them may hinder comprehension.**

0 Vocabulary is inadequate, **too little to judge.**
Almost all sentences contain errors of spelling. **These errors hinder comprehension.**

Grammar

2 A **wide** variety of **grammatical** structures is used.
Almost all grammatical structures are **accurately** and appropriately used.

1 Some variety of grammatical structures is used.
Grammatical structures are **mostly accurate** and/or sometimes inappropriately used.

0 Almost no variety of structures is used.
Grammatical structures are **mostly** inaccurate and/or inappropriately used.

Organisation

2 All ideas are expressed clearly. There is evidence of coherence and/or cohesion throughout the writing.
The writing shows clear paragraphing. Punctuation errors are rare.

1 Most ideas are expressed clearly. There is some evidence of coherence and/or cohesion in the writing.
The writing shows some paragraphing. Punctuation errors may be occasional.

0 Almost all ideas expressed are unclear. There is little/no evidence of coherence and/or cohesion in the writing.
The writing shows little/no paragraphing. Punctuation errors may be frequent.

N.B. The words in bold show the changes to the rating scale.

Appendix C: WRITTEN JUSTIFICATIONS : TASK FULFILMENT

Criterion	Band	Script No.
Task Fulfilment	**0**	**5**

The writing fails to address the writing task. No bullet is covered.
The writing is too short to assess (17 words).
The content is irrelevant to the task
e.g. *When did you came from Slovenia? I woud like to my hause.*

Criterion	Band	Script No.
Task Fulfilment	**1**	**1**

The student tries to address the writing task. However, s/he only partially succeeds in addressing each of the three bullet points, which is why s/he has been awarded 1 point. The teacher/marker has to strain to interpret the writing, e.g. *clouds for clothes, sepreys for surprise* which is another indicator that the task has only been partially fulfilled.

Criterion	Band	Script No.
Task Fulfilment	**2**	**2**

The writing adequately addresses the writing task. However, only two of the bullets are addressed and developed e.g. bullet no.1 [places]:
If you want I can show you my country. [**addresses** the bullet]
We could visit our cities, museums, go on a hike in the mountains and you have to see river Soča, it's beautiful. [**develops** the bullet]

Bullet no.3 [present] has not been adequately addressed. This could be due to: lack of care in reading the task instructions, misunderstanding the requirements of the bullet e.g. the present you wish to get from him (uncle) and not the present you wish to give him, possible lack of comfort on the student's part with the idea of asking for a present.

Appendix D: GLOSSARY

No. of points	TASK FULLFILMENT
3	**FULLY addresses** indicates that the writing: has the required form (e.g. letter/postcard/story), includes the appropriate/required addressee, accomplishes communication goal in every respect. The **content** is **RELEVANT** implies that the student's writing is directly related to the topic set by the task; **All THREE content BULLETS** are addressed and developed.[1]
2	**ADEQUATELY addresses** indicates that the writing: has the required form (e.g. letter/postcard/ story), includes the appropriate/required addressee, mostly accomplishes communication goal. The **content** is **MOSTLY RELEVANT** implies that the student's writing is mostly related to the topic set by the task (80%). This refers to the fact that either **TWO** content bullets are addressed and developed **(1+1)** or **TWO** content bullets are addressed and developed and the third bullet is addressed but not developed **(1+1+1/2)** or **ONE** content bullet is addressed and developed and two other bullets addressed only **(1+1/2+1/2)**.
1	**PARTIALLY addresses** indicates that the writing: has the required form (e.g. letter/postcard/ story), mostly includes the appropriate/required addressee, incompletely accomplishes communication goal. The **content** is **PARTIALLY RELEVANT** implies that the student's writing is not always related to the topic set by the task. This refers to the fact that either: **ONE** content bullet is addressed and developed and another one only addressed **(1+1/2)** or only **ONE** content bullet is addressed and developed **(1)** or **TWO** content bullets are addressed but not developed **(1/2+1/2)** or **THREE** bullets are addressed but NOT developed **(1/2+1/2+1/2)**.
0	**FAILS** to address indicates that the writing: does not have the required form (e.g. letter/postcard/ story), includes an inappropriate/not required addressee, does not accomplish communication goal. The **content** is **IRRELEVANT** implies that the student's writing is **NOT AT ALL** related to the topic set by the task. This means that only **ONE** content bullet is addressed but **NOT** developed **(1/2)** or **NO** content bullet is addressed or developed.

No. of Points	VOCABULARY
3	**RICH** and **ADEQUATE** vocabulary refers to a whole range of vocabulary related to the task. **APPROPRIATE** vocabulary refers to words or expressions that are required by the topic of the set task. **ALMOST NO sentences** – 1 or 2 sentences/clauses.
2	**ADEQUATE** vocabulary is sufficient to the task. **APPROPRIATE** vocabulary refers to words or expressions that are required by the topic of the set task. **A FEW sentences** refers to 3 to 4 sentences.
1	**PARTIALLY ADEQUATE** vocabulary is occasionally insufficient to the task. **NARROW RANGE** refers to basic vocabulary, used to express elementary needs. **REPETITIVE** vocabulary if certain words/expressions are frequently used in the student's writing. **SOME sentences** refers to 5 or 6 sentences/clauses.
0	**INADEQUATE** vocabulary is not sufficient, includes words/expressions that contain errors of spelling or wrong use and therefore cannot be judged by a teacher for their meaning/s. **ALMOST ALL sentences** refers to more than 6 sentences/clauses.

No. of Points	GRAMMAR
2	A **WIDE VARIETY** of grammatical structures refers to **5 or more**.
1	**SOME VARIETY** of grammatical structures refers to **3** or **4**.
0	**ALMOST NO VARIETY** of structures refers to **1** or **2**.

(Bilješke)

[1] If the student has written one clause/sentence about a bullet, s/he has only addressed it and will therefore be awarded with HALF A POINT. However, if the student has addressed the bullet and then developed it in a few more clauses/sentences s/he will be awarded with ONE POINT. E.g. If you want I can show you my country (addresses the bullet); We could visit our cities, museums, go on a hike in the mountains and you have to see the river Soča, it's beautiful (develops the bullet).

VRJEDNOVANJE PISMENIH SASTAVAKA POMOĆU BODOVNE SKALE: STUDIJA SLUČAJA U SLOVENSKOM KONTEKSTU

SAŽETAK:

Vrjednovanje pismenih sastavaka kod testova koji su važni za daljnje školovanje već duže vrijeme privlači pozornost onih koji se bave sastavljanjem testova, vrjednovanjem testova i ostalim istraživačima u području testiranja jezika. Postoji mnogo različitih i neovisnih elemenata koji mogu utjecati na sustav vrjednovanja pismenoga sastava. Bodovne skale koje uključuju ravnomjerne deskriptore koji se nadalje sastoje od specifičnih uvjeta testiranja su od ključne važnosti. Ukoliko opis u skali vrjednovanja nije potpuno jasan te se ne može ravnomjerno primijeniti, rezultati će biti nepouzdani, a bodovanje pismenoga sastavka nepravedno. Da bi se umanjila subjektivnost kod bodovanja pismenoga uratka, mora se uspostaviti procedura sastavljanja i implementiranja mjerila za svako područje u valjanoj bodovnoj ljestvici.

Ključne riječi: vrjednovanje jezika, testiranje jezika, vrjednovanje pisanja, mjerila, bodovna ljestvica, test s visokim ulozima