**Gašper Ilc**
**Veronika Rot Gabrovec**
**Andrej Stopar**
University of Ljubljana*

# RELATING THE SLOVENIAN SECONDARY SCHOOL ENGLISH LANGUAGE NATIONAL EXAMINATIONS TO THE CEFR: FINDINGS AND IMPLICATIONS

## 1 INTRODUCTION

The *Common European Framework of Reference for Languages* (CEFR 2001) had already been successfully adopted and widely used in Europe as a comprehensive set of guidelines and as a valuable reference when, in 2007, the intergovernmental forum on "The *Common European Framework of Reference for Languages* (CEFR) and the development of language policies: challenges and responsibilities", organised by the Council of Europe, clearly stated that each member state is fully responsible to use the document coherently and realistically (Council of Europe 2007). In December 2007, practitioners and academics presented selected case studies at a colloquium in Cambridge, reflecting on their experience of using the draft *Manual,* and reporting about various projects on linking all kinds of examinations to the CEFR (Martyniuk 2010). Only a year later, the Slovenian National Examination Centre launched an ambitious five-year project, relating all the Slovenian national exams in English to the CEFR.

The goal of this article is to present the results of this project, with a particular focus on the two secondary-school-leaving exams: the Vocational Matura and the General Matura. First, the two English exams under investigation are discussed thoroughly, with regard to their construct and format. Second, the results of the five-year alignment project are briefly described. Finally, the general implications and recommendations for both secondary school-leaving exams are presented. The project results and the interpretation of the findings provide clear evidence that the present formats of the exams are not optimal, and suggest that only constant exam validation would guarantee better quality and fairness in the test scores reported in the Slovenian secondary school-leaving language examinations.

---

\*    *Authors' addresses:* Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana, Slovenia. E-mails: gasper.ilc@ff.uni-lj.si, Veronika.RotGabrovec@ff.uni-lj.si, andrej.stopar@ff.uni-lj.si.

## 2  EXAMINATIONS UNDER INVESTIGATION

### 2.1  The General Matura

The General Matura (henceforth: GM) is a national exam sat by test-takers finishing general upper secondary school.[1] The GM serves two purposes. Firstly, it is a school-leaving exam, testing whether the standards of the national curricula for the upper secondary education have been achieved and, secondly, it serves as an entrance exam for higher/university education. Since the GM functions as an achievement, as well as a proficiency exam, the exam administrator and the item writers constantly have to strike a balance between the two types of exams. The fact that the test-takers' chances of entering university solely depend on their exam performance places the GM among the Slovenian national exams with the highest test-taking risk.

The GM consists of five different subjects, three of which are compulsory, and two elective. A foreign language (henceforth, FL) is one of the compulsory subjects. The test-takers can select among six different modern FLs, however, English is selected by the vast majority of the test-takers (approx. 90%). The FL exam can be taken at two levels: the basic and the higher level. In both cases, the basic level is selected by the majority of test-takers, with only 20%–30% selecting the higher level. The GM has two exam dates, one in spring and the other in autumn. Most test-takers (more than 85% of the overall test-takers) sit the exam in spring.

### 2.1.1  GM Exam Construct

The GM for the English Language is specified in the *Subject Testing Catalogue for the General Matura – English* (Ilc et al. 2013, henceforth: GM Catalogue), which is based on the requirements of the *Curriculum for upper secondary school education – English* (Eržen et al. 2008, henceforth: Curriculum). The Curriculum follows the traditional four-language-skill model (reading, listening, speaking, writing), and adopts the CEFR descriptors (CEFR 2001) to specify the skills in greater detail. According to the curriculum, the level expected after completing two years of education is B1, and after four years the expected level is B2. Since the GM is primarily a school-leaving exam, its difficulty level must be aligned with the B2, as specified in the curriculum.

The GM examination for the English Language consists of five subtests that, with the exception of one, correspond to the four-language-skill model, specifically the reading/listening/writing/speaking subtests. In addition, there is a fifth subtest that focuses on the use of English. The exam is administered in two parts. The first part, written, is made up of the reading, listening, writing and the use-of-English subtests, and is administered at the same time nation-wide, and assessed externally by using the same procedures, rules and criteria. It lasts two hours and 50 minutes. The writing subset is assessed by two independent external assessors. The second part, the speaking subtest, is administered by the Matura school committees that use standardised prompts and criteria prepared and specified by the exam provider. The correlation between the two

---

1   Slo. *gimnazija*. It is a four-year secondary education for students aged 14–18.

parts is weak – in the spring of 2013, it was 0.44 for the basic level, and 0.31 for the higher level (Državni izpitni center 2014b: 124–126).[2]

The basic and higher levels of the GM for the English Language share the same structure, however, with the exception of the listening subtest, all the tasks are different.

The reading subtest lasts 35 minutes, and represents 20% of the final grade. The test-takers are required to read two texts, and answer from 16 to 20 accompanying questions. The reading subtest assesses the test-takers' understanding of the main idea (explicit/implicit) and some specific information (explicit/implicit) in authentic texts. The range of text-types in the subtest is varied, and covers a spectrum from formal to semi-formal texts, mostly newspaper articles, fiction passages, etc. The task-types include sentence completion, short answers, multiple choice, information-category/section-summary matching, multiple matching, and gapped text tasks.

The listening subtest lasts 20 minutes, and represents 15% of the final grade. The test-takers are required to listen to two audio tracks, and answer from 14 to 18 accompanying questions. The listening subtest assesses the test-takers' understanding of the main idea (explicit/implicit), and some specific information (explicit/implicit) in authentic texts, as spoken by native speakers of English, with no (or only a slight) dialect variation. The range of text-types in the subtest is varied, and covers a spectrum from formal to informal texts, mostly radio interviews, discussion, etc. The task-types include sentence completion, short answers, multiple choice, information-category matching, and T/F tasks.

The use-of-English subtest lasts 25 minutes, and represents 15% of the final grade. The test-takers are required to complete two language-in-use tasks, together comprising 24 to 30 items. The use-of-English subtest assesses the test-takers' practical knowledge of language use in authentic contexts, e.g. collocations, the use of functional words, verbal forms, word-formation processes, etc. The task-types include one-word gap fill, word-formation, tenses, and multiple choice tasks.

The writing subtest lasts 90 minutes, and represents 30% of the final grade. There are two tasks to complete. Task 1, the prompt-guided task, focuses on functional writing, so the test-takers are required to develop the prompt by using the appropriate text format (letter, e-mail, report, or article) and register (formal or semi-formal). Task 2 is an unguided essay. At the basic level, the test-takers are offered two essay titles, of which they select and develop only one. At the higher level, the test-takers write a literary essay, which is based on two literary works that have been prescribed by the Subject Testing Committee for the Matura. The subtest thus assesses the test-takers' ability to form a written text, using an appropriate text format and register (Task 1), as well as developing an argumentative text (Task 2).

The speaking subtest lasts up to 20 minutes, and represents 20% of the final grade. It consists of three tasks. Task 1 covers a general topic, and is prompted by visual and/or textual cues. Task 2 covers a topic that is tied to test-takers' study programmes, and Task 3 is prompted by a textual cue based on the literary works prescribed by the

---

2    We assume that this may be due to the oral part being administered and rated internally, whereas the rest of the exam is administered and rated externally.

Subject Testing Committee for the Matura. The speaking assesses test-takers' speaking abilities, conversation management, and argument development.

The average difficulty index (ID) for the GM exam across the five components in English is 0.74 (basic level) and 0.76 (higher level), however, the ID varies from subtest to subtest. Typically, the speaking subtest has the highest ID value. Table 1 below shows the average ID for each of the exam subtests administered in spring 2013.

| Subtest | ID (basic level) | ID (higher level) |
|---|---|---|
| Reading | 0.74 | 0.79 |
| Use of English | 0.64 | 0.72 |
| Listening | 0.57 | 0.68 |
| Writing | 0.76 | 0.81 |
| Speaking | 0.83 | 0.96 |

Table 1: The average ID for the GM subtests in 2013 (Državni izpitni center 2014b: 156)

## 2.2 The Vocational Matura

The Vocational Matura (henceforth VM) is a national school-leaving examination giving the test-taker a technical/vocational-technical education. As stipulated in the Slovenian Matura Examination Act (Uradni list RS 2007), by taking the exam, the test-takers demonstrate their achievement of the standards required by the curricula of technical secondary schools, vocational-technical schools, and vocational courses. Successful VM test-takers can continue their studies at the tertiary level, as students of vocational and other colleges, without any further requirements.[3] The VM consists of four subjects: Slovenian, a general elective (Mathematics or a FL), a technical-theoretical and a practical subject. English is one of three FLs offered to the VM test-takers.

VM test-takers come from a variety of different educational institutions (vocational schools, technical schools, general secondary schools, and adult education courses), so they represent a heterogeneous population. They differ from each other with respect to course content, course load, forms of instruction (full-time studies versus part-time vocational courses), and age (in 2013 around 10% of the test-takers were older than thirty; cf. Državni izpitni center 2014a: 21).

The VM has three exam dates: in spring, autumn and winter. The majority of test-takers sit the exam in spring. In 2013, the number of spring test-takers was 8,960: 57% chose Mathematics as their third VM subject, and 43% a FL. English is, by far, the most popular FL at the VM – 91% of all FL test-takers chose it in 2013.

---

3    If the test-takers wish to enrol into a university-level programme, they are required to sit for a general matura subject, in addition to taking the four VM subjects. According to the VM 2013 Annual Report (Državni izpitni center 2014), in 2013 about 15% of all VM test-takers also sat a general matura exam. Fewer than 70% passed.

### 2.2.1    *VM Exam Construct*

The VM for the English Language is specified in the *Subject Testing Catalogue for the Vocational Matura 2014 – English* (Andrin et al. 2012; henceforth: VM Catalogue), which follows the requirements of the curriculum, as described in the *Catalogue of Knowledge Standards for FL1*[4] (Andrin et al. 2011; henceforth: VM Curriculum). The exam simultaneously assesses the level of acquired knowledge, the test-takers' proficiency in English and, partly, their knowledge of language for specific purposes. Thus, it exhibits characteristics of achievement tests, proficiency tests and specific purposes tests (cf. Alderson et al. 1995: 12).

The VM Curriculum explicitly states that the teaching and learning of a FL is "based on the guidelines of the Common European Framework of Reference for Languages" (Andrin et al. 2011: 4). The CEFR is referenced a number of times, e.g. in the sections on receptive and productive skills, text-types, and language competence. The levels A1 and A2 are referred to and described as synonymous with "a lower level of complexity", and the level B1 as signifying "a higher level of complexity". In the section describing the minimal standards, the document states (Andrin et al. 2011: 16) that "[t]he expected level of knowledge of the first foreign language at the end of vocational education is A2 in accordance with the Common European Framework." The minimal standards for technical secondary schools and vocational-technical schools, on the other hand, contain a different recommendation stating that "[t]he expected level […] is B1 […]" (Andrin et al. 2011: 19).

The VM Catalogue is always aligned with the VM Curriculum. The current exam scheme in the catalogue is as follows (cf. Andrin et al. 2012: 8). The written part lasts 120 minutes and consists of a 60-minute reading comprehension paper and a 60-minute written communication paper, each representing 30% of the final grade. The oral part lasts 20 minutes and represents 40% of the final grade. The correlation between the written and the oral parts is moderately high: in 2013 it was 0.61.

The text-types in the reading comprehension paper run the gamut from informal to formal. Often found in the exam are newspaper articles, descriptions of procedures, descriptions of people, interviews and instructions for use. The reading task-types (usually four per paper) include short-answer tasks, matching tasks, multiple-choice questions, gapped-texts and T/F tasks, and others. The reading subtest aims to assess the test-takers' understanding of the main message, and some specific information in authentic texts. Moreover, the test-takers should also be able to identify the main points of a text, understand sequences of events, understand emotions and points of view, and be able to use dictionaries (cf. Andrin et al. 2012: 6).

The written communication section lists target text-types, such as letters, guided or unguided compositions based on verbal or visual prompts, descriptions, journal entries, and others. The typical writing subtest includes two tasks: a short, 70-word composition, based on a visual or verbal prompt, and a long, 150-word guided or unguided composition. The subtest assesses the test-takers' ability to form and link sentences, paragraphs and messages; to organize information (opinions, interests and feelings); to form different

---

4    The same curriculum document is used for both English and German (as FL1).

types of texts; to distinguish between formal and informal texts; to spell correctly using dictionaries; and to use various writing strategies (Andrin et al. 2012: 6).

The oral part consists of three tasks: one is general in content, and two are tied to test-takers' study programmes (i.e. they are ESP-based tasks). Test-takers are expected to participate in role-plays, to summarize and comment on a short written prompt, to talk about a topic indicated by a visual prompt, to explain and discuss a topic related to their vocation/profession, and the like. According to the VM Catalogue (Andrin et al. 2012: 7), the test-takers should be able to correctly pronounce words; use appropriate intonation patterns; interact; introduce themselves and others; talk about their experiences, work and vocation; report and describe; summarize and mediate information; express opinions and feelings; and use appropriate communication strategies.

The VM for the English Language does not include a listening subtest. However, according to the VM Catalogue (Andrin et al. 2012: 7), at the oral exam, the test-takers show their abilities to understand (by listening) the gist of the message and react to it; to understand specific information; and to recognize circumstances such as the mood of the interlocutor and intercultural peculiarities.

The facility values for the 2013 spring exam date are presented in Table 2.

| Subtest | ID |
|---------|------|
| Reading | 0.75 |
| Writing | 0.75 |
| Speaking | 0.84 |

Table 2: The average ID for the VM subtests in 2013 (Državni izpitni center 2014a: 106)


## 3  PROJECT BACKGROUND

In 2008, the Slovenian National Examination Centre launched a CEFR-relating project that encompassed all national exams in English. Most of these exams are based on national curricula which, in turn, serve as exam constructs. Consequently, any curricular change is directly mirrored in the exams themselves. One of the project objectives was also to determine to what extent these exams are comparable with international exams that had already been aligned with the CEFR (Bitenc Peharc and Tratnik 2014). The first stage (2008–2010) of the relating project centred on receptive, and the second (2011–2013) on productive, skills.

As stipulated in the project report (Bitenc Peharc and Tratnik 2014), the relating team (henceforth the panellists) used the procedures as recommended by the Manual (CEFR 2001) and the Council of Europe (c.f. Council of Europe 2009; Council of Europe 2004). Moreover, the panellists also relied on the EALTA Guidelines for Good Practice in Language Testing and Assessment (2010), and other guidelines for, or reports of, good practices in testing/assessment, as developed by experts in the field (Alderson 1995; Bachman 1990; Bachman and Palmer 1996; Cizek 2001; Cizek and Bunch 2007; Hughes 2002; Madsen 1983; McNamara 2000; Weir 1990 etc.)

The panellists first underwent the familiarization stage, during which they were acquainted with the CEFR and its descriptors, and trained in the CEFR relating procedures. During the second, specification stage, the panellists analysed and described the exam in terms of the Manual (Council of Europe 2009), which consequently required a thorough examination of the subject curricula, the subject testing catalogues, and other specifications. The two exams under investigation herein, the General Matura and the Vocational Matura in English, were placed at the levels B2 and B1, respectively. The third stage, the standardisation stage, closely adhered to the procedures specified in the Manual (Council of Europe 2009), and in the Reference Supplement (Council of Europe 2004). For aligning the receptive skills, the standard-setting methods employed were the Basket and the Extended Tucker-Angoff Methods (Council of Europe 2009: 61–66, 75–77), whereas for aligning the productive skills, the benchmarking method (Council of Europe 2009: 35 et pass.) was applied. The reason for combining the Basket and the Angoff Methods for setting the cut scores for the receptive skills lies in the established finding that applying different methods leads to different cut scores (cf. Kaftandijeva 2010: 131). To minimize the panellists' subjectivity,[5] the alignment was constantly monitored by continuous familiarization procedures and discussions among panellists. This procedure was followed as each skill was addressed, and before each round of judgements: first, after levels had been assigned to the items and, second, after the statistics on item difficulty, assigned levels, and correlations among panellists were revealed (by way of EXCEL spreadsheets). With regard to the setting of the cut-scores, the panellists first set the cut-scores using the Basket and the Angoff methods separately, and then averaged the two results into the final cut-score.

The last stage of the project, the validation stage, mostly involved finalising and documenting the linking procedures. As highlighted in the final report (Bitenc Peharc and Tratnik 2014: 28), high correlations were observed between the judgements and statistical data (-0.8–-0.97 for the Basket Method, 0.73–0.99 for the Angoff Method, and 0.86–0.99 for the benchmarking method).

## 3.1    Project Results – the GM

The table below presents the results of the alignment project, as provided by the final project report (Bitenc Peharc and Tratnik 2014). The results show that the majority of items of the GM are aligned with the CEFR level B2; therefore, we can claim that the GM targets at the CEFR level B2. The compound cut-score for the basic level is set at 77%, and at 71% for the higher level.

---

5    The panellists' judgments could have been influenced either by their different occupational backgrounds (primary school teachers, secondary school teachers, university teachers, and testing professionals), or by their shared cultural and educational background (all are Slovenian nationals educated in Slovenia).

| Subtest | Basic Level | | | Higher Level | | |
|---|---|---|---|---|---|---|
| | Items* (%) | Cut-score (%) | Population** (%) | Items* (%) | Cut-score (%) | Population** (%) |
| Reading | B1: 48 B2: 50 C1: 2 | 80 | 44 | B1: 42.5 B2: 52.4 C1: 5 | 80 | 85 |
| Listening | B1: 41 B2: 53 C1: 6 | 79 | 55 | B1: 41 B2: 53 C1: 6 | 79 | 55 |
| Use of English | B1: 20 B2: 65 C1: 15 | 70 | 21 | B1: 26 B2: 58 C1: 16 | 72 | 68 |
| Writing: Task 1 | - | 80 | 40 | - | 70 | 95 |
| Writing: Task 2 | - | 75 | 35 | - | 56 | 91 |
| Speaking | - | 80 | 68 | - | 80 | 97 |

Table 3: The GM alignment results

* CEFR levels as determined by the panellists.

** The cut-score was applied to the test-takers who took the investigated GM subtest. The percentage indicates the test-takers who achieved the targeted level.

## 3.2 Project Results – the VM

The table below presents the results of the alignment project, as provided by the final project report (Bitenc Peharc and Tratnik 2014). The results show that the majority of items of the VM are aligned with the CEFR level B1, therefore, we can claim that the VM targets at CEFR level B1. The compound cut-score is set at 71%. At the spring 2013 exam date, 2324 (74%) out of 3128 test-takers achieved this level, whereas the 51% pass mark that was set by the Subject Testing Committee was achieved by 97% of the test-takers.

| Subtest | Items* (%) | Cut-score (%) | Population** (%) |
|---|---|---|---|
| Reading | A2: 25 B1: 53 B2: 22 | 70 | 67 |
| Writing: short composition | - | 70 | 75 (for the combined writing communication cut-score of 67%) |
| Writing: long composition | - | 65 | |
| Speaking | - | 75 | 80 |

Table 4: The VM cut-scores

* CEFR levels as determined by the panellists.

** The cut-score was applied to the test-takers who took the investigated GM subtest. The percentage indicates the test-takers who achieved the targeted level.

## 4 IMPLICATIONS AND RECOMMENDATIONS

Even though the alignment project was conducted in accordance with the methodological guidelines (cf. Section 3), there are several issues to consider. First, it is questionable to what extent the shared educational/cultural background of the panellists affected their judgements. North and Jones (2009), for instance, claim that a shared educational/national/cultural background can have a significant impact on the CEFR-relating process. Pižorn (2009) also reports on the initial problems the panellists faced at the early stages of the alignment process. However, as partial external evaluations of the alignment project (Ilc and Stopar 2014) suggest, these factors had an insignificant effect on their judgements.

Second, it is not clear whether the cut-scores, as stated by Bitenc Peharc and Tratnik (2014) are valid and applicable for the interpretation of exam results from different exam dates. It has to be pointed out that the project results for each individual subtest rely on the analysis of only one exam booklet. Since the GM for the English Language is not piloted or pretested, it is difficult to assume that different exam booklets display exactly the same level of difficulty. The relative comparability of different exam booklets depends solely on the better judgement of the item-writers and the Subject Committee, which is not supported by any (statistical) analysis prior to the administration of the exam. Therefore, it is methodologically problematic to apply the same cut-scores to different exam booklets.

Third, the test provider should also encourage and conduct a constant exam validation, and linkage of the exam booklets. To assure test validity and comparable levels of difficulty, the test provider should also introduce pretesting, piloting, and anchor items. The item-writers should be well-acquainted with the CEFR descriptors, and provide a better ratio between the targeted CEFR level questions, and questions belonging to the two adjoining levels.

Leaving these objections aside, it should be acknowledged that the proposed cut-scores are comparable with the cut-scores of international exams in English targeted at the same CEFR level,[6] with the exception of the low cut-score of 56% for the GM Task 2 in the writing subtest on the higher level (cf. Table 2 in 3.1) and, perhaps less so, for the VM long composition with the cut-score of 65% (cf. Table 4 in 3.2). We should stress that these observations are based mostly on raw scores. To establish a more valid comparison between these exams, we should not rely merely on raw scores, but rather on a more detailed comparative study investigating the exam constructs, methods, text-types, tasks, context, cognitive processes involved, etc.

### 4.1 Specific Recommendations for the GM

Even if it is accepted, for the sake of argument, that the project results are valid, and that the established cut-scores can be applied to different exam booklets, at least two fundamental issues arise. First, the results show that a significant number of subtest questions

---

6    E.g. the test-provider for the First Certificate in English (FCE) indicates that test-takers achieving
     Grade B (75%–79%) display "the range of good achievement" (Cambridge English 2013).

were judged as B1 and C1 items, which are either below or above the level targeted by the Curriculum. While this may be typical of proficiency/diagnostic exams (Alderson 2000; Bachman and Palmer 1996), it is not desirable in the case of achievement exams, because any question above the targeted level may impede the test-takers' results. This becomes even more problematic when we take into consideration the fact that, at the spring 2013 exam date (Table 5), only 39% of test-takers achieved the targeted B2 level overall – i.e. if we apply the established cut-scores (cf. Table 3 in 3.1). This means that 61% of test-takers, who were expected to be at the level B2 (cf. the curriculum aims in 2.1.1), not only struggled with C1 questions, but also found the B2 questions challenging. On the other hand, a relatively high proportion of B1 questions, which may indeed be the only suitable questions for the 61% of the test-takers, challenges the well-established status of the GM as a B2 exam (cf. Table 3 in 3.1), and questions the GM exam validity, at least in terms of its construct (i.e. B2 targeting Curriculum).

| Subtest (spring 2013) | Basic level* (%) | Higher level* (%) |
|---|---|---|
| Reading | 39 | 60 |
| Listening | 11 | 29 |
| Use of English | 36 | 58 |
| Writing – Task 1 | 62 | 98 |
| Writing – Task 2 | 52 | 98 |
| Speaking | 71 | 98 |
| Overall exam | 39 | 91 |

Table 5: The application of the established cut-scores (Bitenc Peharc and Tratnik 2014) to the spring 2013 exam date results
* The percentage indicates the test-takers who achieved the targeted level.

Secondly, the fact that more than 50% of the test-takers (61% of the basic level test-takers at the spring 2013 exam date) fail to achieve B2 calls for an in-depth analysis to be provided by the curriculum designers, as well as educators. In particular, they should address the question of why such a relatively high proportion of test-takers fail to achieve the curricular goals.

The results of the alignment project discussed above show that the present format of the GM is not optimal. In order for the test scores of this high-stakes exam to be valid, the following points should be considered.

(i) **Reading and listening subtests.** The current format of the reading and listening subtests differs noticeably from many well-established international English language examinations, in terms of the number of questions and text-type varieties. Firstly, the reading subtest should consist of a variety of texts – descriptive, narrative, expository, argumentative, and instructive (Alderson 2000: 127) – and of different task-types. To assure the validity of the subtest, it should include from 30 to 40 different questions (see Alderson 2000: 294). Secondly,

the listening subtest should comprise a variety of audio tracks, i.e. at least three different audio tracks, as well as task-types. The number of questions should be increased to at least 20, preferably to 30, questions. In the present format, if the test-takers fail to answer one question out of 14–18 questions in the listening subtests (c.f. 2.1.1), they will automatically lose from five to eight percent overall. Taking the cut-score of 79% into consideration, this means that a B2 test-taker may only fail to answer two questions.

(ii) **Exam format.** Presently, the exam can be taken at two levels. It would, however, be more appropriate if the exam were administered on one level only, since this would allow greater differentiation and discrimination (Bitenc Peharc and Tratnik 2014: 38). Additionally, to identify the B1 test-takers, two sets of cut-scores should be determined (B1 and B2).[7] Alternatively, if the present format of the exam on two levels is preserved, it is of vital importance that the test provider selects appropriately calibrated items that accurately reflect the difference in difficulty between the two levels.[8]

(iii) **Speaking subtest.** Given the poor correlations between the external and internal parts of the exam (cf. 2.1.1), the questions and the subtest administration procedure should be more standardised. In particular, the interlocutors and assessors should undergo the same standardisation procedures as the assessors of the writing subtest. If would be advisable if the assessor(s) were appointed externally.

## 4.2    Specific Recommendations for the VM

Two issues are likely to affect the interpretation of project findings, with regard to the content of the VM for the English Language. Firstly, the alignment shows that a relatively high proportion of the items in the reading subtest were judged as B2 items, which is above the CEFR level targeted by the VM Curriculum, and may have a negative impact on the success of the weaker test-takers. The opposite could be claimed for the A2 items included in the exam, however, these are more easily justifiable, since the CEFR A2 descriptors are also included in the VM Curriculum (admittedly, mostly as a target for vocational schools). Secondly, the VM does not include a listening comprehension subtest, which raises doubts about the validity of the exam as a whole (see Bachman and Palmer 1996; Brown 1996 a.o.).

The noticeable gap between the number of B1 test-takers and the number of passing test-takers should also be addressed, with regard to the VM rating procedures. Specifically, the 40% oral part is based on the guidelines included in the VM Catalogue, but is still prepared, administered and assessed internally, at the vocational/technical educational institutions. Furthermore, the subtests for reading comprehen-

---

7    It is noteworthy that these, and the previous, recommendations can be achieved only if the entire time allotment of the GM changes.

8    Only this step will enable the test-provider to avoid the situation in which a subtest of a lower level contains more difficult questions than a subtest from a higher level, as is also shown by the project results (cf. Bitenc Peharc and Tratnik 2014).

sion and writing communication (representing 60% of the overall grade) are prepared externally, however, their rating is also conducted internally. Although the exam papers are anonymized, the assessors are likely to recognize their students' handwriting, especially since the number of test-takers at a particular school can be very small. Thus, we should observe that some deficiencies of the VM may affect the results of the exam and, consequently, contribute to the existence of the gap between the number of B1 test-takers and the number of passing test-takers that has been exposed by the alignment project.

Nevertheless, if we assume that the project results are valid, that the VM subject testing committee consistently produces tasks on the same level of difficulty, and that the deficiencies of the exam have a negligible impact on the results, the explanation for the gap between the B1 test-takers (the 71% cut-score) and the successful test-takers (the 51% pass-score) may lie in the curriculum. On the one hand, the targeted CEFR level B1 could be overly-demanding for VM test-takers – since they are a heterogeneous population, this is not necessarily an unexpected finding. Consequently, it might be justifiable to set two cut-scores (A2 and B1) for the VM population. Alternatively, if the level B1 remains the curricular goal for the VM population, the current VM pass score of 51% is set too low. Since introducing the 71% B1 cut-score as the pass score would result in a mere 74% of the spring 2013 test-takers passing the VM for the English Language, the curriculum designers and educational institutions will have to evaluate why such a substantial part of the VM population obtains a passing school grade (which allows them to sign up for the VM exam), but fails to achieve the prescribed CEFR level.

In light of the above, we propose the following.

(i)   **Construct validity**. To properly align the VM for the English Language with its construct, the VM Curriculum, a listening subtest should be introduced (see also Bitenc Peharc and Tratnik 2014: 38). It is worth noting that all other Slovenian national exams in English contain this subtest.

(ii)   **Assessment procedures**. To ensure demonstrable objectivity, an external rating system (similar to the one implemented for the GM) should be enacted, that includes standardised and externally-prepared oral exam tasks; external assessors for the speaking subtest; external and completely anonymous assessment of the writing subtest; and machine-aided scoring of the reading comprehension subtest.

(iii) **Curriculum.** Curriculum designers should reconsider the curriculum aims in terms of realistic achievable goals, utilizing the data obtained from the alignment project.

## 5   CONCLUSION

The results of the five-year project, conducted by the Slovenian National Examination Centre, show that – though neither of the Matura examinations is yet optimal

– they both contain the majority of items that target at the CEFR levels, as specified by the their constructs, i.e. the National Curricula (the level B2 for the GM, and the level B1 for the VM). Thus, we can assume that the two exams are aligned with their respective CEFR levels.

Although this means that the levels of the exams reflect the aims, objectives and standards as set by the respective National Curricula, there are still some unsettling matters to be discussed. One of them is the dual character of the GM: the GM has always been both an achievement (school-leaving) and a proficiency (university entrance) exam. The project results show that a significant number of subtest questions were judged either below, or above, the level targeted by the Curriculum which, at least to some extent, questions the GM exam validity as an achievement exam.

Another matter that needs addressing is the correlation between the written and oral parts of both exams. The correlation between the (externally assessed) written and the (internally assessed) oral part of the GM is weak, and while the correlation between the parts of the VM seems moderately high, one has to keep in mind that the rating of all the VM papers is conducted internally. This leads to the conclusion that the quality of the link between Slovenian secondary school language exams and the CEFR proficiency levels will only be truly validated once *all* the examiners, the external and the internal ones, and indeed *all* the EFL secondary school teachers are familiar with its proficiency levels. In particular, the responsibility to administer valid and fair tests to all the test-takers nationwide lies not only with the Slovenian National Examination Centre and, of course, the policy makers, but also with each individual secondary school examination board, and all the English teachers involved in the teaching and testing process. This might also help avoid the rather high number of students who complete the school year successfully, and then unfortunately fail to achieve the prescribed CEFR level. Only then will the aptitude for taking responsibility for "making coherent, realistic use of the CEFR" (Council of Europe 2007) be genuinely shown.

## References

ALDERSON, Charles J. (1995) *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

ALDERSON, Charles J. (2000) *Assessing Reading.* Cambridge: Cambridge University Press.

ALDERSON, Charles J./Caroline CLAPHAM/Diane WALL (1995) *Language test instruction and evaluation.* Cambridge/New York/Melbourne/Madrid/Cape Town/Singapore/Sao Paulo: Cambridge University Press.

ANDRIN, Alenka/Mira HLADNIK/Mateja PETROVIČ/Tanja STEYER/Andrej STOPAR (2012) *Predmetni izpitni katalog za poklicno maturo – angleščina.* Ljubljana: Državni izpitni center. 8 May 2014. http://www.ric.si/mma/P-ANG-2014%20 ISSN/2012092610051308/.

ANDRIN, Alenka/Helena KOZAR/Suzana REHBERGER/Susanne VOLČANŠEK/Mihaela U. ZAKOŠEK (2011) *Katalog znanja. Prvi tuji jezik – angleščina/nemščina.* 8 May 2014. http://portal.mss.edus.si/msswww/programi2011/programi/SPI/KZ-IK/KZ_TJ1_2011.doc.

BACHMAN, Lyle F. (1990) *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

BACHMAN, Lyle F./Adrian S. PALMER (1996) *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford: Oxford University Press.

BITENC PEHARC, Suzana/Alenka TRATNIK (2014) *Umestitev nacionalnih izpitov iz angleščine v skupni evropski okvir. Zaključno poročilo o izvedbi projekta.* Ljubljana: Državni izpitni center.

BROWN, James D. (1996) *Testing in Language Programs.* Upper Saddle River, NJ: Prentice Hall Regents.

CAMBRIDGE ENGLISH (2013) *Cambridge English First, First Certificate in English (FCE), CEFR Level B2.* 8 February 2014. http://www.cambridgeenglish.org/images/25091-fce-level-b2-document.pdf.

CIZEK, Gregory J. (ed.) (2001) *Setting Perfomance Standards, Concepts, Methods, and Perspectives.* Mahwah/New Jersey/London: Lawrence Erlbaum Associates, Publishers.

CIZEK, Gregory J./Michael B. BUNCH (2007) *Standard Setting, A Guide to Establishing and Evaluating Performance Standards on Tests.* Thousand Oaks/London/New Delhi: SAGE Publications.

*Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (2001). Cambridge: Cambridge University Press.

COUNCIL OF EUROPE (2004) *A Reference Supplement.* Strasbourg: Language Policy Division.

COUNCIL OF EUROPE (2007) *The Common European Framework of Reference for Languages* (CEFR) *and the development of language policies: challenges and responsibilities*. Strasbourg: Intergovernmental policy forum.

COUNCIL OF EUROPE (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual.* Strasbourg: Language Policy Division.

DRŽAVNI IZPITNI CENTER (2009) *Splošna matura, predmeti, angleščina, jesenski rok 2009.* 11 August 2013. http://www.ric.si/splosna_matura/predmeti/anglescina/#jesenski_rok_2009.

DRŽAVNI IZPITNI CENTER (2014a) *Letno maturitetno poročilo o poklicni maturi 2013.* Ljubljana: Državni izpitni center. 9 May 2014. http://www.ric.si/mma/Letno%20porocilo%20za%20poklicno%20maturo%202013/2014052913535193/.

DRŽAVNI IZPITNI CENTER (2014b) *Letno maturitetno poročilo o splošni maturi 2013.* Ljubljana: Državni izpitni center. 22 June 2014. http://www.ric.si/splosna_matura/statisticni_podatki/.

ERŽEN, Vineta/Alenka BUDIHNA/Berta KOGOJ/Blanka KLOBUČAR/Boštjan VRE-CL/Ingrid ZUPANC-BREČKO/Jelka NAPAST/Karmen PIŽORN/Katica SEMEC PEVEC (2008) *Učni načrt. Angleščina. Gimnazija; Splošna, klasična, strokovna gimnazija: obvezni ali izbirni predmet in matura (420 ur)*. Ljubljana: Ministrstvo za šolstvo in šport, Zavod RS za šolstvo. 22 June 2014. http://portal.mss.edus.si/mss-www/programi2012/programi/media/pdf/un_gimnazija/un_anglescina_gimn.pdf.

HUGHES, Arthur (2002) *Testing for Language Teachers*. Cambridge: Cambridge University Press.

ILC, Gašper/Alenka KETIŠ/Aleksandra KOMADINA/Ana LIKAR/Simona MEGLIČ/Irena ZORKO NOVAK (2013) *Predmetni izpitni katalog za splošno maturo – angleščina.* 22 June 2014. http://www.ric.si/splosna_matura/predmeti/anglescina/.

ILC, Gašper/Andrej STOPAR (2014) "Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English." *Language testing* [online first]. "http://dx.doi.org/10.1177/0265532214562098" \t "doi" 10.1177/0265532214562098.

KAFTANDIJEVA, Felianka (2010) *Methods for Setting Cut Scores in Criterion-referenced Achievement Tests. A comparative analysis of six recent methods with an application to tests of reading in EFL*. Arnhem: Cito, EALTA. 5 February 2014. http://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf.

MADSEN, Harold S. (1983) *Techniques in Testing*. Oxford: Oxford University Press.

MARTYNIUK, Waldemar (ed.) (2010) *Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual (Studies in Language Testing)*. Cambridge: Cambridge University Press.

McNAMARA, Tim (2000) *Language Testing*. Oxford: Oxford University Press.

NORTH, Brian/Neil JONES (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages ... Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling.* Strasbourg: Language Policy Division. 11 January 2014. https://www.coe.int/t/dg4/linguistic/Manual%20-%20Extra%20Material%20-%20proofread%20-%20FINAL.pdf.

PAPAGEORGIOU, Spiros (2010) "Investigating the decision-making process of standard setting participants." *Language Testing* 27/2, 261–282.

PIŽORN, Karmen (2009) "Designing Proficiency Levels for English for Primary and Secondary School Students and the Impact of the CEFR." In: N. Figueras/J. Noijons (eds), *Linking to the CEFR levels: Research perspectives*. Arnhem: Cito, EALTA, 87–102. 14 February 2014. http://www.ealta.eu.org/documents/resources/Research_Colloquium_report.pdf.

URADNI LIST RS (2007 January 5) *Zakon o maturi, uradno prečiščeno besedilo (ZMat-UPB1)*. Ljubljana. 12 May 2014. http://www.uradni-list.si/1/objava.jsp?stevilka=3&urlid=20071.

WEIR, Cyril J. (1990) *Communicative Language Testing*. New York/London/Toronto/Sydney/Tokyo/Singapore: Prentice Hall.

Summary
RELATING THE SLOVENIAN SECONDARY SCHOOL ENGLISH LANGUAGE
NATIONAL EXAMINATIONS TO THE CEFR: FINDINGS AND IMPLICATIONS

The present paper draws on the report of a five-year project that aligned the Slovenian national exams in English to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). Discussed here are the key findings of the relating project, carried out by the National Examination Centre, a central institution for external assessment in Slovenia, for the following exams: the Vocational Matura (for technical secondary schools, vocational-technical schools, and vocational courses) and the General Matura (for general secondary education programmes). The focus of the paper is on the interpretation of the findings of the project, the significance of relating the aforementioned exams to the CEFR, the implications of project results for future language test development and, most importantly, the impact of the findings on the development of secondary school-level English education programmes in Slovenia.

**Key words:** CEFR, alignment, national examinations in English, testing, curriculum.

Povzetek
UMEŠČANJE IZPITOV IZ ANGLEŠKEGA JEZIKA NA SPLOŠNI IN POKLICNI
MATURI V SKUPNI EVROPSKI JEZIKOVNI OKVIR (SEJO): UGOTOVITVE IN
NJIHOV VPLIV

Prispevek oriše petletno umeščanje slovenskih nacionalnih izpitov iz angleškega jezika v Skupni evropski jezikovni okvir (SEJO) in predstavi ključne ugotovitve projekta, ki ga je izvedel Državni izpitni center, osrednja institucija za zunanje preverjanje znanja jezikov v Sloveniji. Predstavljena sta izpit iz angleščine na poklicni maturi (za srednje tehniško in drugo strokovno izobraževanje, poklicno-tehniško izobraževanje in poklicne tečaje) in izpit iz angleščine na splošni maturi (za gimnazije). Prispevek je osredinjen na interpretacijo rezultatov projekta umeščanja, na pomen umeščanja obeh izpitov v CEFR, na vpliv rezultatov projekta na razvoj jezikovnih testov v prihodnje, in – kar je najpomembneje – na vpliv ugotovitev umeščanja na razvoj srednješolskih programov v Sloveniji.

**Ključne besede:** SEJO, umeščanje, nacionalni izpiti iz angleščine, preverjanje, učni načrt.