EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 5, 19–32

CONSTRAINTS OF MEASURING LANGUAGE PROFICIENCY IN ESTONIA: THE NATIONAL EXAMINATION IN THE ENGLISH LANGUAGE

Ene Alas, Suliko Liiv

Abstract. The current article gives an overview of the development and problems related to the advancement of the national examination in English in Estonia over a ten-year period, starting from its launch in 1997. The process started in 1994, after Estonia regained its independence, and proceeded from the need to standardise both foreign language instruction and evaluation. The national examination gave the Ministry of Education, schools, teachers and students an opportunity to adequately assess language proficiency, as well as compare students and schools. On the other hand, universities and businesses obtained a tool to make admission/recruitment decisions. The article discusses the principles of the national examination construction, its specification, structural alterations over time, the task types implemented to measure particular language skills, marking procedures, exam results and exam evaluation.

Keywords: test validity, test reliability, test specifications, rater reliability, washback effect

Introduction

It was in 1994 that the first attempts were made to systematically start to follow the principles that had been established in the western tradition of language testing for some time and had been outlined in the works of Underhill (1987), Weir (1988), Hughes (1989), Bachman (1990), Alderson, Clapham and Wall (their then unpublished manuscript of the seminal 1995 testing book) to name but a few. Language testing research and test development in the west were by that time independent, indispensable parts of foreign language instruction and evaluation and Estonia, with its newly regained political, economic and cultural independence, was in a hurry to learn from the western experience and implement the principles in the English

language evaluation practice here. Evaluating the foreign language testing situation in Estonia prior to establishing a national exam, Tallinn University professor of English Suliko Liiv, who has long been a foreign language teaching and evaluation policy maker in Estonia, asserts that "...there was no unified school-leaving examination in English, teachers had a great deal of freedom in compiling, administering and marking the tests. Each school compiled their own tests for final exams and the result was that the tasks varied a great deal and the results of the exams in different schools were not comparable and tended to be subjective." (Liiv 2002: 51–52)

It was primarily this problem that drove the Ministry of Education in 1994, shortly after Estonia regained its independence and was starting to align its teaching and evaluation practices with those followed in the west, to look for "a common yardstick...in order to make meaningful comparisons" (Hughes 2003: 4), to give teachers a common standard that would allow them to measure their students against and to allow students to compare their own proficiency against, to give the Ministry of Education a tool to make comparisons between schools and allow the schools and universities to use the same tool for gatekeeping purposes. So it was clear from the start that what was attempted was going to be a high stakes test.

The need to create an instrument that would be utilised to measure the language ability across Estonia prompted the then Ministry of Education to put together a working group that started to develop the first pilot tests. The effort to launch a national test for upper-secondary/high-school/gymnasium graduates was not restricted to the English language only, but involved all languages taught in Estonia (Estonian, Russian, English, German, French) and also sciences. A lot of general training for test developers at the start of the project was conducted to all subject specialists together, but to date, all subjects-specific national examination development groups are working fairly independently. In order to fully concentrate on the development of a national qualification evaluation system, the National Examination and Qualification Centre (NEQC) opened in 1997 that currently oversees national examination development[1] among other things and has, as one of its chief responsibilities, to guarantee timely and professional national examination management. The scope of this article will not allow us to make comparisons with other subject areas, thus the discussion will be restricted to the English language national examination development only.

The English language national exam today

The design and development of the English language national exam proceeds from the Ministry of Education and Science regulation of January 23, 2001 no. 18 "Õpitulemuste välishindamise põhimõtted, riigieksamitööde, põhikooli eksamitööde ja üleriigiliste tasemetööde koostamise, hindamise ja tulemuste hindamise alused"[2] (Regulation 2001). The regulation specifies the purposes of the national exam as follows:

- To evaluate the attainment of the educational goals outlined in the basic and gymnasium curricula;
- To give schools and teachers an opportunity to compare the results of their students to those achieved by other students in the country;

[1] The analysis of the national examinations of the English language is yearly published by NEQC, see NE 1997 – NE 2007.

[2] "Principles of external evaluation of study results, standards for compilation, evaluation and results' analyses for national examinations, basic school final papers and state standard tests."

- To steer the educational process through the content and form of national examinations;
- To link consecutive educational levels and stages;
- Through external marking, to give feedback to all stakeholders and to allow planning and execution of changes in the national curriculum, textbooks, in-service training of teachers and allow development in the respective areas.[3]

As can be seen, the purpose of the national exam has in broad terms remained similar to its initial envisaged purpose. Consequently, what the exam developers have to constantly be aware of is the enormous washback effect in terms of teaching and testing practices at school and its impact on the stakeholders. "Stakeholders would include the test designers, teachers, students, score users, governments or any other individual or group that has an interest in how the scores are used and whether they are useful for a given context" (Fulcher, Davidson 2007: 14). The impact of the exam can be illustrated with just a few examples. Out of 59 specialities admitting students to Tallinn University BA level studies in 2008, 24 specified the foreign language national examination result as being of criterial importance during the admission procedure. The number of students who have chosen English as their national (graduation) exam over the years and consequently perceived it to be of relevance for their subsequent career choices can be seen in the Table 1.

Table 1. Number of participants in the English national exam over the years

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Examinees	9280	8769	9258	9461	8488	9311	9431	9099	9415	9590	9696

Considering that the average overall number of gymnasium graduates in Estonia is slightly above 14 000, it can be seen that generally speaking, about 75 per cent of the school leavers choose English as one of their graduation exams.[4]

The development of the national exam proceeds according to specifications that are derived from the national curriculum on the one hand, and from Year 12 Handbook, on the other. The national curriculum specifies study goals, competencies and skills to be acquired within a specified amount of time (cf. Curriculum 2002). The study goals in the national curriculum are outlined very broadly. An example of the kind of specifications one can find there is the specification concerning gymnasium graduates' oral proficiency: a student demonstrates oral proficiency through "employing the correct foreign language intonation, rhythm and stress; being able to converse within the specified topical range by presenting and supporting his/her point of view; by knowing the communication etiquette and being able to use it; by being able to communicate in the foreign language both directly and by telephone; by being able to exchange information, ask questions and express their position on social problems and events; and by resorting to compensatory strategies in communication if necessary" (ibid.). The topic areas specified are the following: **I as an individual** among other individuals, my special features, abilities, preferences, strengths and weaknesses; **family and home**, marriage and family, roles in the family, rights and obligations, family budget; **friends**, relations between friends, social problems; **environment, Estonia, the world**, nature and nature protec-

[3] NEQC, www.ekk.edu.ee (05.09.2008).
[4] www.ekk.edu.ee (2.08.2008).

tion, natural resources, climate, town and country, urbanisation, Estonian government, economy, cultural traditions, international relations; **English-speaking countries**, governments, culture, international relations; **everyday activities**, healthy ways of life, nutrition, communication in service situations, help during emergencies; **study and work**, the system of education, opportunities for education in Estonia and English-speaking countries, study skills and exam techniques, work and unemployment, technological advancement; **hobbies and culture**, sports events, cultural figures, advertising, information society and its problems (ibid.). The language level to be achieved by the end of gymnasium studies in the English language is B2 in all subskills (reading, writing, speaking and listening) as defined in the Common European Framework of Reference for Languages (CEF 2001).

The curriculum thus specifies the content of the examination in very broad terms. A much more concrete exam specifications can be found in the Year 12 Handbook. The first of its kind was published in 1995 and was subsequently edited numerous times as the exam developed. The handbook describes each sub-skill (writing, listening, reading and language structures) paper and the speaking test in detail, gives examples of possible text types and task types, provides sample answers, tips for the student, and marking scales for the subjectively evaluated sections of the exam (writing and speaking) (cf. Jõul et al. 2005).

Relying on the specifications, the next task for the national examination development team is to compile a test that would first and foremost be valid and reliable, i.e. test the proficiency that it claims to test and do so irrespective of the conditions and occasions of testing. Each national examination paper is a team effort, which draws its tasks from the effort of a number of item writers, who have been trained to write items to test a particular skill. This procedure, too, has evolved over the years. If at first, the whole exam development team was involved with all the exam tasks, the work now is divided between skill teams. It is the skill team leader who receives the items or complete tasks from item-writers, assembles the items into tasks and submits them to the English language chief specialist. The items/tasks go through moderation carried out by independent consultants and are then all piloted usually among the 11th formers in different schools to evaluate their effectiveness. The schools that are chosen represent the whole spectrum of schools whose students sit the national exam, i.e. town schools, country schools, Russian schools, Estonian schools, etc. Piloting of test items involves both statistical and qualitative evaluation. As Hughes (2003: 65) points out, the statistical analysis at this point will "reveal qualities (such as reliability) of the test as a whole and of individual items (for example, how difficult they are, how well they discriminate between stronger and weaker candidates)". The qualitative analysis, on the other hand is carried out "in order to discover misinterpretations, unanticipated but possibly correct responses, and any other indicators of faulty items". Once satisfactory pilot results are achieved (which usually means more changes to the tasks and sometimes dropping of the tasks as unsuitable), the chief specialist puts together two final versions of the national exam (Variant A and B), both ideally of equal quality. Before the finalisation, the exam versions are trialled and proof-read by native speakers of English. This procedure of test development has been followed with very few changes starting from 1997 when the first national test was put together.

The Table 2 below illustrates the structure of the current national exam in the English language, specifying the number of tasks in each section, the maximum number of points available for that section and the time allotted for the completion of the section.

Table 2. National examination structure

	Skill	Tasks	Maximum points	Time (min.)
1	Writing	2	20	80
2	Listening	3	20	35
3	Reading	4	20	50
4	Language Structures	4	20	40
5	Speaking	2	20	13–16

The time given for each section has generally remained the same over the years with two exceptions. In 2001, the time for the listening section was extended from 30 minutes for 35 minutes and in 2006, the time for the writing section was raised from 75 minutes to 80 minutes. Tasks 1–4 are completed consecutively on the same day, with the speaking test taken on the following day. Compared to other skill papers, the speaking test allows the examiner some freedom as to the time within which the test has to be completed. This is done in order to consider the idiosyncrasies of the examinees, allowing for varying rates of response and speech speed.

Task types

The **writing** paper has two tasks, the first of which is a letter and the second task is either an essay or a report. The expected length for a letter up until 2006 was specified as between 80 and 120 words. In order to avoid awarding similar points for exam responses of substantially differing lengths (e.g. one student writing 80 words and scoring maximum points and another student writing 120 words and also scoring maximum points) the requirement was changed as of 2007 where all the examinees are expected to write 120 words and are penalised if the response is significantly shorter. Another change in this task involved the genre. If the Year 12 Handbook in 2005 still specified the expected text types as "form filling, formal letters, instructions, notes and messages, postcards and personal letters" (Jõul et al. 2005: 14), then, relying on the national curriculum guidelines and the CEF B2 level writing (CEF 2001: 61–62), the tasks that are effectively set in this part of the exam are semi-formal or formal letters of different genre (e.g. inquiry, apology, complaint, protest, etc.), all other genres (writing a postcard, leaving a message, etc.) are expected to have been mastered at a lower level. The second writing task can currently be either an essay or a report. Due to marking constraints, the story, which used to be a potential task type on this level, was excluded from the list as of 2007. In fact, although as a task type, the story features in specifications prior to 2007, it never appeared as an actual task in the national examination. The required length for the second writing task (essay/report) was set at 200 words in 2007. Here, too, a range (from 150 to 200) was allowed prior to that, which potentially may have given rise to unfair test scores.

The **listening** comprehension paper has three tasks that employ text types such as public announcements, interviews and conversations between two or

more people, mini-lectures, radio programmes, etc. Every consecutive task has an increased level of difficulty, which is decided by to the pilot stage results. The tasks vary from one exam to the next but are either yes/no, multiple choice or short answer questions, matching tasks, ordering tasks, completing tasks or information transfer tasks. A huge and persistent challenge with the listening comprehension test is quality control of the recordings – finding suitable non-copyrighted texts, choosing speakers for the original recordings (the accent, the speed, the tone of voice, etc. of the speakers), making decisions about the background noise.

The **reading** paper, similarly to the other papers, derives its topics form the national curriculum. The paper contains three texts that are each followed by one or two tasks. The texts originate from brochures, leaflets, forms, letters, instructions, advertisements, fiction, reference books, journals and magazines, dictionaries, etc. Typical task types are multiple choice and true/false questions, matching of titles and paragraphs, matching words with definitions, interview questions with responses, inserting deleted sentences, ordering paragraphs. The task type that causes perennial debate within the paper is the true/false/no information task that places huge demands on the item writers to create items which clearly belong in just one of the given categories (true or false or no information) and is not interpretable in more than one way.

The **language structures'** paper focuses most specifically on the grammatical accuracy and appropriacy of the English language use. It is this part of the language competence that has been specified in the most detail in the national curriculum (for the list of grammatical requirements for a upper-secondary school/gymnasium/ high school graduate see for example Jõul et al. 2005, appendix E, 131–133). The challenge for the test writers is to achieve appropriate coverage of the specifications. If well designed, this section allows "checking the students' knowledge within a fairly short amount of time of very different language structures, also those that in a daily language feature less frequently" (NE 2001: 19). The grammar structures are checked within complete, connected texts. It is not sufficient to be familiar with particular grammatical items only to complete this section of the exam successfully. It is necessary to know how to implement the grammatical knowledge within a particular text. Thus a successful completion of tasks also requires attentive reading of the tasks on top of grammar knowledge. It is here that we notice that dividing language tests into skill tests is somewhat arbitrary in that by testing one skill we are inadvertently also testing another (in this case, while testing structures, we are also testing the reading skill).

The **speaking** test takes place on a day following the written papers (depending on the size of the school, it may take between 1 and 3 days to administer the speaking test to all the students who have registered for it) and currently requires the examinee to complete two tasks: a monologue and a (two-participant) role-play. The prompt for the monologue has gone through a thorough process of evolution, proceeding from a picture (until 2001), to a quote (2001–2002), a short article (2003–2007), and currently, a controversial statement (as of 2008). The main reason for substituting short articles as prompts was the attempt to reduce the amount of reading in the speaking test. As can be seen from the discussion above, the national examination already has a fairly heavy bias on testing reading (the reading paper, and the language structures' paper). The new format allowed the examinee to focus on displaying his/her speaking skills without depending on the

reading-comprehension first. This part of the national exam has been updated most recently for the purposes of higher reliability. Both tasks of the exam are scripted, i.e. the interviewer has to follow a prescribed format for the interview and is not allowed to improvise or deviate from the wording of the script. Improvisation may lead him/her to ask questions of varying levels of difficulty from different examinees, leading to unequal treatment and potentially unfair marking. Following a script will ensure equal conditions for all examinees, irrespective of the examination day, the time of the day, the order of the examinees and the fatigue level or the personal characteristics of the interviewer.

Marking procedures

Both objective and subjective marking have been implemented with the national examination in the English language from the very start. Listening, reading and language structures' papers have always been marked objectively, relying on the answer key for each item. Providing the answer key is a simultaneous process to item writing but also continues during the piloting stage, which invariably produces occasional acceptable but previously overlooked answers. Once the answer key is complete, no judgement is required on the part of the marker. A special case are the tasks in the listening paper that require students to fill gaps or provide short answers, and consequently issues of correct spelling come into play. Thus here a complete answer key cannot be prepared prior to test administration. To ensure uniform marking, a standardisation meeting is called after the examination paper has been administered and a random sample of about one hundred papers is taken to determine the extent of spelling diversion accepted as correct. In principle, no "points for errors of grammar or spelling [are deducted], provided that it is clear that the correct response was intended" (Hughes 2003: 170). It is, however necessary to determine where the line of clarity runs. When the respective decisions are made, the marking proper will proceed according to the key compiled.

Writing and speaking sections of the national exam are subjectively marked, i.e. teams of raters are trained either to rate the students' writing papers or their performance during the speaking test. In writing, the raters have generally relied on two different marking scales – one for letters and another for the essays and reports. With the number of point available for a particular paper fixed – 20 points as a sum total for both tasks – the major concern while developing the marking scales has always been what to reward within the skill. The marking scale for letters has moved from awarding points for task completion, letter format and language (1999) to evaluating task completion, vocabulary and register, and grammar and spelling (2001), to task completion, letter format and language (until 2006) and task completion and language (as of 2007). It is also interesting to note that until the 2007 scale, specific sub-skills had been weighted differently. An example is the 1999 scale, where for task completion the students could get the maximum of 2 points, but for vocabulary and register and for grammar and spelling a maximum of 3 points. In the 2006 letter scale, task completion and format both earned the writer a maximum of two points, but the language criterion was evaluated on the scale of 0 to 4. This type of marking may inadvertently lay the classroom teaching emphasis on language (i.e. grammar and vocabulary) and overlook other facets of

writing, such as content and organisation, thus disadvantaging the student, should he/she move to such language contexts where the aforementioned qualities of writing are required. For a more detailed discussion of the 2007 national examination writing scales see Alas et al. 2006. All writing papers are marked by two raters and in case of a disagreement of 4 points or more in the evaluation results, a third rater is called in for a final decision.

The marking of speaking has undergone substantial changes, too. The challenges for the rating scale development are similar to those with the writing scales, i.e. which criteria to select for evaluation. Here, too, the scale has moved from a full scale for all the criteria selected in 1999, to an unequal number of points allocated for different criteria (as of 2001) back to a full scale starting from 2007. The current marking scale evaluates the students' performance from the point of view of four criteria – communication, vocabulary, grammar, and pronunciation and fluency. For a full discussion of the 2007 speaking scale see Alas 2007. The students' oral performance is rated by an independent examiner during the oral exam. The examiner does not participate in the interview, which takes place between the student and the interviewer, but only rates the student's performance relying on the marking scale.

Exam results

All five exam sections are equally weighted – the maximum number of points that can be awarded for each section is 20, thus the maximum number of points the examinee can receive for the whole exam is one hundred. Below, an attempt will be made to draw some conclusions from a decade of the English language national examination administration in Estonia. The comparison and analysis will rely on the national examination 1997–2007 results. The Table 3 below shows the average scores of the student who have taken the national exam in the English language over the years along with the standard deviation i.e. the "average amount that each student's score deviates from the mean" (Alderson et al. 1995: 294), the maximum number of points gained and the minimum scored during a particular test.

Table 3. Examinees and their mean score

Year	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**
Examinees	9280	8769	9258	9461	8488	9311	9431	9099	9415	9590	9696
Average	64.6	58.8	61.8	64.1	64.9	66.6	63.99	66.6	71.9	64,4	68.8
Std*	17.7	19.9	19.9	19.7	18.8	17.8	16.9	16.7	16.0	16.1	16.0
Max	99	99	100	99	99	100	100	100	100	99	99
Minimum	8	0	0	0	5	0	0	1	1	11	5

* Std = standard deviation

Looking at the average scores, which is just one of the very many statistical data derived from each year's test result, it can be observed that with two exceptions the mean score has remained relatively stable during the decade. It is only in 1998, that the average score has dropped to 58.8 points, which may indicate a relatively more difficult test compared to the others. In 2005, however, the average score suddenly shoots to 71.9, which in turn points at a somewhat easier national exam.

With these two exceptions, the examination development team has managed to produce fairly uniform exams.

It is also worthwhile comparing the average scores awarded for particular skills within the exams. The Table 4 makes comparisons between the average scores calculated over the years (1998–2007) for a particular skill as well as juxtaposes it with the averages for the other four sections of the test.

Table 4. Overview of mean scores for skills (1998–2007)

Year	Writing	Listening	Reading	Structures	Speaking
1998	12.2	10.1	10.7	10.4	15.6
1999	12.4	11.2	10.9	11.8	15.7
2000	12.3	11.6	13.3	9.9	15.6
2001	11.3	14.7	12.2	11.1	14.7
2002	11.6	13.2	14.7	11.9	15.5
2003	11.5	11.9	13.5	11.0	15.8
2004	13.4	12.0	13.7	11.5	16.1
2005	13.3	12.7	15.3	13.1	16.4
2006	12.9	11.3	11.9	12.1	16.6
2007	13.1	13.1	12.5	13.1	16.9

Comparing the results across the board, it can be seen that while writing, listening, reading and language structures seem to correlate fairly well with one other, the average score for speaking is significantly higher every year. If these scores are reliable, then the students' speaking skills are for some reason significantly higher than all the other skills. Given that successful speaking presupposes good vocabulary, a good command of grammatical structures and the ability to interact with the interlocutor (hearing, understanding and responding to what is said, i.e. listening skills), the result is somewhat dubious from the point of reliability Another factor that may skew the results is the fact that although the schools are urged to record the examinees, and the examinees are urged to request recording of their oral interviews (without a recording the student cannot appeal against their interview result), this is not general practice. Thus all the interviews are marked by just one rater whose judgement is hardly ever monitored, which may lead to a tendency to inflate the score in an attempt to compensate for possible lower scores in other sections of the test.

The students' average results have already been discussed above. It would, however be interesting to look at different groups of students. The Table 5 shows the average results of male and female students from the time when such comparative data are available.

Table 5. Mean score of boys and girls

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007
Boys	60.3	61.0	63.3	66.2	63.3	65.5	71.4	65.2	69.5
Girls	62.8	63.6	64.6	66.9	64.4	67.3	72.3	63.8	68.3

The Table 5 shows that with two exceptions (2006 and 2007), the girls results have generally been higher, which may indicate a slightly better language competence level of girls, but could also be an indicator that the exam items have been con-

structed so that they are more accessible to the female population of test takers. From the raters' comments it seems to transpire that girls are generally better at completing writing and speaking tasks while boys are more successful in listening, reading and language structures.

Another point of comparison is the medium of instruction at school. Estonia has both Estonian and Russian language schools, where the primary language of instruction is Estonian or Russian, respectively. The same exam is available as a national exam for both school types. The average results of the students can be seen in the Table 6.

Table 6. Mean score of Russian and Estonian students

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Estonian	66.8	61.2	64.4	64.6	65.8	68.3	65.6	67.5	74.1	66.3	70.7
Russian	59.2	51.5	53.5	55.6	59.1	61.8	59.3	64.2	65.2	57.8	68.8

A study of the results demonstrates a significantly higher average every year of the students studying in the Estonian language schools. The difference may be explained by the fact that while most of the Estonian-speaking test takers have studied English as an A-language (the first foreign language that the students start studying), the vast majority of the Russian-speaking students taking the test have started studying English as a B-language (the second foreign language, which begins two years later). Thus by the time the examinees take the exam, the Russian students would have studied English for a shorter period of time.

Exam evaluation

All English language national exams are post-validated through a battery of statistical data dealing with item analysis, looking at "(1) the degree to which the item discriminates among individuals of different levels of ability (the discrimination parameter); (2) the level of difficulty of the item (the 'difficulty parameter') and (3) the probability that an individual of low ability can answer the item correctly (the 'pseudo-chance' or 'guessing' parameter)" (Bachman 1990: 204). Although the data is available for the English national exams over the years, the discussion of it is no within the scope of this article. As a step in test validation and development, however, post-validation is of utmost importance as it gives test developers feedback on the quality of their work and the trustworthiness of the test results.

Problems

In spite of the huge strides made in the field of language proficiency testing, there are problems, some of which (like the reliability of spoken language testing) have been discussed above, that remain.

The national development team is constantly looking for more item writers, which would hopefully considerably contribute to a more varied and higher quality items. It could also speed up the test construction process and avoid last minute decision-making.

There is the concern of compiling two equally valid and reliable variants for the English national exam every year, where the first variant of the test is taken by the vast majority of the test whereas the second variant is taken by very few (e.g. in 2006, 9552 people took variant A and 38 people took variant B). The resources that go into the development of both variants, however, are equal and seem somewhat wasted with so few students taking variant B.

The third concern involves test security and pertains to the level of information given to the teachers and students about the national examination without actually giving away the particular test items, tasks and questions. The existing test construction procedure that relies on a great number of item writers supervised by skills team leaders who in turn relinquish the tasks to independent consultants and subject specialist alongside with general training for teachers in exam techniques, marking scales' implementation and testing practices hopefully guarantees secure (and thus valid and reliable) tests on the one hand and a reduced level of teachers' national exam elated anxiety on the other, but it needs honing.

Conclusion

Estonia has been involved in professional test-construction for over a decade and that has given the Estonian education system an enormous amount of experience.

The English language national exam is well-established. It is the most widely taken, locally constructed, nation-wide foreign language proficiency exam in Estonia which is comparable to other national foreign language exams in Europe. The exam writers are guided by the standards adopted by the Council of Europe, expressed in the Common European Framework of Reference for Languages. The Estonian national curriculum specifies B2 as the language level required in English from the Estonian gymnasium graduates. The curriculum outlines in very broad terms the different CEF levels but to date, the levels have not been sufficiently elaborated. The national exam in English is a B2 level exam insofar as it proceeds from the CEF principles and tries to align its tasks and language content with other English language proficiency exams that have the B2 status (e.g. FCE).

Estonia has become a member of international testing organisations like Association of Language Testers in Europe (ALTE), European Association for Language Testing and Assessment (EALTA), etc., proceeding in the test construction from their codes of practice.

Test writers and developers know how to construct valid, reliable tests and administer them professionally. Test construction follows internationally established guidelines and practices of test specification, item writing, piloting, test administration and statistical analysis. Test construction has had a washback effect on the language teaching practices at school, with the teachers being trained in the best practices of how to teach and test a particular skill, how to choose a textbook and supplement it so that it would benefit the student most. All past tests are on file and available for students and teachers to learn from on the National Examination and Qualification Centre home page.

There is a greater awareness among educators of concerns that surround testing. Testing has become a specific subject taught in the teacher education courses. Raters are systematically trained to make expert decisions about student writing

skills and oral performances. Semi-annual workshops and conferences are held to familiarise teachers with the national test development issues and give them feedback on past practices. Besides learning from the European practice of language testing, testing experts from Estonia share their expertise of test construction in Estonia at international conferences.

Abbreviations

ALTE – Association of Language Testers in Europe
CEF – Common European Framework of Reference for Languages
EALTA – European Association for Language Testing and Assessment
FCE – First Certificate in English (B2 level Cambridge test)
NEQC – The National Examination and Qualification Centre (= Riiklik Eksami- ja Kvalifikatsioonikeskus)

References

Alas, Ene 2007. Developing the national examination in the English language. – Open!, 32, 2–5.

Alderson, J. Charles; Clapham, Caroline; Wall, Dianne 1995. Language Test Construction and Evaluation. Cambridge Language Teaching Library. Cambridge: Cambridge University Press.

Bachman, Lyle F. 1990. Fundamental Considerations in Language Testing. Oxford Applied Linguistics. Oxford: Oxford University Press.

CEF 2001 = Council of Europe 2001. Common European Framework of Reference for Languages: Learning, Teaching and Assessment 2001. Cambridge: Cambridge University Press. http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf (10.03.2009).

Curriculum 2002 = Põhikooli ja gümnaasiumi riiklik õppekava 2002. Riigi Teataja, I, nr 20, Tallinn.

Fulcher, Glenn; Davidson, Fred 2007. Language Testing and Assessment. An Advanced Resource Book. London, New York: Routledge Applied Linguistics.

Hughes, Arthur 1989. Testing for Language Teachers. Cambridge Language Teaching Library. Cambridge: Cambridge University Press.

Hughes, Arthur 2003. Testing for Language Teachers. 2nd ed. Cambridge Language Teaching Library. Cambridge: Cambridge University Press.

Jõul, Mare; Lätt, Viive; Mere, Kristi; Sass, Eve; Türk, Ülle; Vilu, Maila 2005. Year 12 Handbook. Tallinn: Argo.

Liiv, Suliko 2002. Foreign language competence and testing. – Suliko Liiv (Ed.). Perspectives on English and American Language and Literature. Tallinn: Tallinna Pedagoogikaülikooli Kirjastus, 51–59.

NE 1997 = Inglise keel. Riigieksam 1997. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 1998 = Inglise keel. Riigieksam 1998. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 1999 = Inglise keel. Riigieksam 1999. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2000 = Inglise keel. Riigieksam 2000. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2001 = Inglise keel. Riigieksam 2001. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2002 = Inglise keel. Riigieksam 2002. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2003 = Inglise keel. Riigieksam 2003. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2004 = Inglise keel. Riigieksam 2004. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.
NE 2005 = Inglise keel. Riigieksam 2005. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.
NE 2006 = Inglise keel. Riigieksam 2006. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.
NE 2007 = Inglise keel. Riigieksam 2007. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.
Regulation 2001 = Õpitulemuste välishindamise põhimõtted, riigieksamitööde, põhikooli lõpueksamitööde ja üleriigiliste tasemetööde koostamise, hindamise ja tulemuste analüüsi alused. Haridusministri määrus nr 18, 23.1.2001. Tallinn.
Underhill, Nic 1987. Testing Spoken Language: A Handbook of Oral Testing Techniques. Cambridge Handbook for Language Teachers. Cambridge: Cambridge University Press.
Weir, Cyril J. 1988. Communicative Language Testing. Exeter: University of Exeter.

Ene Alas (Tallinna Ülikool). Teadushuvid on keeletestimine, testide koostamine ja nende kvaliteedi hindamine, õpetajakoolitus, õppekirjanduse hindamine.
ene.alas@tlu.ee

Suliko Liiv (Tallinna Ülikool). Uurimisvaldkonnad on kontrastiivuuringud, kultuuridevaheline suhtluspädevus, keelepoliitika, võõrkeelte õpetamise metoodika.
liiv@tlu.ee

KEELEPÄDEVUSE MÕÕTMISEST EESTIS: INGLISE KEELE RIIGIEKSAM

Ene Alas, Suliko Liiv
Tallinna Ülikool

Artikkel annab ülevaate inglise keele riigieksami arengust Eestis ja sellega kaasnenud probleemidest kümne aasta jooksul alates eksami loomisest. Riigieksami arendamise protsess sai alguse 1994. aastal pärast Eesti taasiseseisvumist ja tulenes vajadusest standardiseerida keeleõpetus ja keeletestimine Eestis, et nii haridusministeeriumil, koolidel, õpetajatel kui õpilastel oleks võimalik keeleoskust adekvaatselt hinnata, tulemusi nii individuaalselt kui ka kooliti võrrelda. Teiselt poolt vajasid ülikoolid ja muud asutused usaldusväärset teavet keeleoskuse taseme kohta, et ühtlustada vastuvõtu/töölevõtu põhimõtteid. Artiklis kirjeldatakse inglise keele riigieksami koostamise põhimõtteid ja eksami eristuskirja, eksami ülesehitust ja selles aja jooksul tehtud muudatusi, osaoskuste testides kasutatavaid ülesandetüüpe, hindamise põhimõtteid, hindamisskaalasid ja neis aja jooksul toimunud muudatusi, samuti eksami tulemusi.

Võtmesõnad: testi valiidsus, testi reliaablus, testi eristuskiri, hindajate reliaablus, testi tagasimõju