A Validation Study of the National Assessment Instruments for Young English Language Learners in Norway and Slovenia

Karmen Pižorn*[1] and Eli Moe[2]

This article is a validation study of two national large-scale tests that measure the language proficiency of 11/12 year-old English learners in Norway and Slovenia. Following the example of Alderson and Banerjee (2008), the authors of the article have employed the EALTA guidelines for good practice to validate the tests, and to formulate major recommendations for improvement of both assessment instruments, where feasible (Alderson & Banerjee, 2008). The results of the validation study show that both national tests in English seem to fulfil most of the EALTA guidelines for good practice, although a few issues related to the test construct and test design procedures need to be re-assessed, and some changes may be required.

Keywords: national test, validation, English, EALTA guidelines for good practice

1 *Correspomding author. Faculty of Education, University of Ljubljana, Slovenia
karmen.pizorn@pef.uni-lj.si
2 University of Bergen, Norway

Študija ugotavljanja veljavnosti dveh nacionalnih preizkusov znanja iz angleščine kot tujega jezika pri mlajših učencih na Norveškem in v Sloveniji

Karmen Pižorn* in Eli Moe

Prispevek predstavlja študijo ugotavljanja veljavnosti dveh nacionalnih preizkusov znanja iz angleščine kot tujega jezika pri učencih, starih 11/12 let, na Norveškem in v Sloveniji. Po vzoru Alderson in Banerjee (2008) sta avtorici prispevka uporabili EALTA-smernice za dobro prakso za preverjanje veljavnosti preizkusov ter oblikovanje ključnih in izvedljivih predlogov za izboljšanje obeh sistemov vrednotenja (Alderson in Banerjee, 2008). Izsledki raziskave kažejo, da oba sistema vrednotenja dosegata zahteve EALTA-smernic za dobro prakso, kljub temu pa bi bilo treba nekaj elementov, povezanih z izvedbo in s postopkom preverjanja znanja, ponovno oceniti, saj bi bile mogoče potrebne določene spremembe.

Ključne besede: nacionalno preverjanje znanja, veljavnost, angleščina, EALTA-smernice za dobro prakso

In 2000, Rea-Dickins highlighted the extent to which the teaching of foreign languages was no longer restricted to secondary education (Rea-Dickins, 2000). Twelve years later, a number of countries throughout the world had moved initial foreign language teaching from secondary to primary school or even to the pre-school level (Commission, 2008; Graddol, 2006). Due to this intense activity associated with the teaching of foreign languages at an ever earlier age, the previous two decades have seen an increased focus on the research and development agenda for assessment at this level (Low, Brown, Johnstone, & Pirrie,1995; Mckay, Hudson, & Sapuppo, 1994; Edelenbos & Johnstone,1996; Breen et al.,1997; Leung & Teasdale, 1997; McKay, 2000, 2006; Brumen, Cagran, & Rixon, 2009). While many studies have examined the issues and implications arising from formative assessment of the young foreign language learner, only a few have been concerned with the issues related to the assessment of young foreign language learners using large-scale tests (McKay, 2006; Eurydice, 2009).

The objective of this paper is to validate two national foreign language tests for young learners (11/12 year olds) in two European countries by using the EALTA guidelines for good practice. This validation study demonstrates to what extent the tests fulfil their intended use and what benefits they may bring to the stakeholders involved in the foreign language learning and teaching process, as well as which limitations have to be considered carefully and openly.

Background Information on Educational and Assessment Contexts in Norway and Slovenia

Children in both countries start school when they are six. Compulsory education is for ten years in Norway and nine in Slovenia. In Norway, students change schools between primary and lower-secondary levels, while in Slovenia the schools are single-structured. In both countries, English is one of the core subjects, taught from Year 1 in Norway and from Year 4 in Slovenia.

National tests in English are administered for Year 5 and 8 students in Norway, and for Year 6 and 9 students in Slovenia. The objects of this study are the tests for Year 5 students in Norway and for Year 6 in Slovenia. While all Norwegian Year 5 students sit for the national test in English in September every year, the test is optional for Slovene Year 6 students. However, more than 80% of the entire school-aged population has taken it in the previous nine years. The test in Slovenia is paper-based and comprises listening, reading and writing skills and two tasks assessing vocabulary and grammar, while the test in Norway is computerized, and tests reading, vocabulary and grammar. None of the tests assess students' speaking skills.

Students in both countries are familiarised with the test formats through the test specification documents, old test papers or sample tasks publicly available online, and by their foreign language teachers, who are recommended to inform students about such issues as the testing procedures, test methods and test goals.

The Overview of the national tests in English at primary school in Norway and Slovenia

Table 1: Slovenia: The structure of the national test in English for Year 6 students

Language skill tested	Number of test tasks	Number of items per test task	Number of points	% of total	Purpose
Listening	2	6	6	25	To test students' listening comprehension skills (skimming, scanning, listening for gist etc.)
		6	6		
Use of Language	2	6	6	29	To test use of vocabulary in context
		8	8		
Reading	2	6	6	25	To test students' reading comprehension skills (skimming, scanning, reading for gist etc.)
		6	6		
Writing	1	10	10	21	To test students' writing skills with short guided texts

Table 2: Norway: The structure of the national test in English for Year 5 students from 2012

Language skill tested	Number of test tasks	Number of items per test task	Number of points per item	% of total	Purpose
Reading	14	1–6	1	48	To test students' reading comprehension skills (finding information and understanding main points)
		A total of 24 items	Maximum 24 points		
Vocabulary	21	1	1	42	To test comprehension of vocabulary in context
		A total of 21 items	Maximum 21 points		
Grammar	1	5	1	10	To test grammar in context
		A total of 5 items	Maximum 5 points		

Tables 1 and 2 show that the test structures differ with regard to the skills they assess as well as to the number of items for each skill/language component. This means that while only a small portion of the curriculum goals for Year 5 students are actually tested in Norway, the test in Slovenia covers approximately three quarters of the goals. In both countries, therefore, teachers are responsible for formatively assessing non-tested skills and other language elements.

Examining the validity of national assessments in English at primary school in Slovenia and Norway

The method

Since increasing numbers of primary young foreign language learners are being included in national foreign language assessments for a variety of reasons (monitoring, accountability, diagnosis, etc.), it is crucial that such tests be critically evaluated, preferably while they are being developed, but at least when being implemented. For this purpose, various language testing associations have developed guidelines and codes of good practice. Despite the abundance of testing guidelines, there is little research to document how these are followed and maintained in the course of a practical test development and its use (De Jong & Zheng, 2011). In this article, the authors have used the European Association for Language Testing and Assessment (EALTA) Guidelines for Good Practice in Language Testing and Assessment (EALTA, 2006) as an evaluation instrument. Although the EALTA guidelines address different audiences, the two national tests were examined in relation to test designers' audience. Alderson and Banerjee (2008) argued that guidelines, such as the EALTA guidelines could be used to "frame a validity study" (Alderson, 2010, p. 63) and thus offer recommendations for improvement to test developers and other stakeholders. De Jong and Zheng (2011) suggested that, although the guidelines are very useful as a checklist during the process of test development, they are probably not the ultimate tool for assessing the quality of the test.

The test development processes of the Norwegian and Slovene National tests in English were examined against the three features out of seven critical aspects as defined by the EALTA guidelines, and already used in the two validation studies (Alderson & Banerjee, 2008; De Jong & Zheng, 2011).

The three features:

(1) Test Purpose and Specification;

(2) Test Design and Item Writing and

(3) Quality Control and Test Analyses

were selected according to their importance in developing young language learners' tests and the test data available at the time of writing the article.

The following sections are organized in the order of the three aforementioned features. Each feature is presented by raising a number of questions, the answers to which are given first for Slovenia and then for Norway.

The EALTA Guidelines for Good Practice as a Framework for Validating the National Tests of English at Primary Level in Slovenia and Norway

Test Purpose and Specification

How clearly is/are test purpose(s) specified?

Slovenia: The role of the National English test for Year 6 students is formative, having a focus on recognising the needs of individual students, and providing teachers with additional information about their students' achievements. Another objective is to measure whether the curriculum goals have been met.

Norway: The main aim of the national test of English for Year 5 students is to provide information on the students' basic skills in English on a national level. A secondary aim is to use the results as a basis for improving pupils' English skills.

How is potential test misuse addressed?

Slovenia: To avoid potential misuse of the test, detailed information on how to appropriately interpret and use the test scores is provided in documents available on the National Testing Centre website.[3] The annual report of the test results and the live papers with the key for each test task are made available on the day of the assessment. Another very useful document is the so-called Quartile Analysis, which is accompanied by a comprehensive analysis of the test items falling in each quartile. It provides teachers with a more detailed qualitative description of the students' achievements, which helps them to interpret the scores appropriately and to provide detailed and contextualised feedback. The test is optional and low-stake and the school results are not published. However, on the forum hosted by the National Institute of Education,[4] language teachers expressed their concern about the pressure they were under from the head teachers and parents, who demand better and better results on national tests, without considering differences in the social and intellectual backgrounds of students. This was related to the low results their students had achieved in the test in May 2012, when the test difficulty dropped from .73 to .59.

3 http://www.ric.si/

4 http://skupnost.sio.si/mod/forum/discuss.php?d=27516

Norway: The Year 5 national test in English is administered to all pupils in that year. This is done through a national test administration system, which means that the test cannot be administered to pupils for whom it was not developed. In addition, the tests are low-stake for the pupils, since no decisions regarding their future education or lives are made on the basis of the results. Nevertheless, since aggregated test results are published, the tests often become high-stake for schools and teachers. Since the first national tests were administered in 2004, it has become clear that at some schools more or less all Year 5 students take the test, while at others a certain percentage of pupils are "absent". The fact that schools have different practices with regard to test attendance is something the education authorities now plan to investigate. This is related to the fact that many stakeholders have a negative attitude towards the publication of results; it is one thing that school owners, i.e. local authorities, have access to test results, but it is quite a different matter when local and national newspapers rank schools on the basis of test results. The Norwegian Ministry of Education and the Norwegian Directorate for Education and Training do not encourage the publication of test results in the press, but since official aggregated test results are public by law, there is nothing to prevent newspapers from "doing their worst" in this matter. Unless the law is changed, this practice will continue.

Are all stakeholders specifically identified?

Slovenia: The test stakeholders include students of Year 6 throughout Slovenia, English language teachers, primary school head teachers, curriculum experts and policy makers. The test specifications and the annual report include different kinds of information that may be used by individual stakeholders. For example, the information on pupil performance, combined with other data provided, is intended for the head teachers and policy makers specifically; while the detailed descriptions of individual test tasks and pupil performance, as well as the quartile analysis, may be of great value to the language teachers. There is also a short, reader-friendly brochure with information about the national assessment in primary schools; it is intended to help parents and pupils to understand the main aims of the tests. However, there is no document specifically for parents, with explanations and advice on what to do if their child's score on the test is very low or very high.

Norway: All of the stakeholders, and the responsibilities of the various stakeholders, are clearly identified and defined. National and local educational authorities are responsible for circulating information about the test, examining the test results on different levels and, if necessary, taking action and making changes on the basis of the information collected. Teachers and schools are responsible

for ensuring that the subjects are taught in a way that makes it possible for pupils to achieve curriculum goals. In addition, local school authorities, head teachers and teachers have specified tasks to perform before, during and after test administration. Parents are informed about the purpose and content of the test, as well as of their child's test result. Test quality requirements are specified, and the test developers have to develop tests in accordance with these and to demonstrate this compliance using statistical analysis, and in official piloting and test reports.

Are there test specifications?

Slovenia: The test specifications do not exist as one comprehensive document but take the form of a number of documents: (1) The Test Structure, which provides a detailed description of the test, primarily intended for the test designers and language teachers; (2) The Information for Students and Parents; (3) The Administration Guidelines for the National Assessment in Primary School, which describes the administration of the test in detail and is mainly intended for the head teachers and teachers; (4) The Quartile Analysis of the students achievements, which is a very thorough description of the test items in relation to the pupils' achievements; and (5) the sample tasks and old test papers with answer keys, and assessment criteria from 2005 onwards.

Norway: Test specifications exist and are available in the teachers' guidelines, which provide information on test purpose, test construct, test takers, test format, item formats, number of items and scoring procedures are specified, and links are provided to sample tasks and the previous years' test. Some of this information is also included in an information brochure for parents.

Are the specifications for the various audiences differentiated?

Slovenia: The test specifications are mainly intended for teachers, test designers and, to some extent, for researchers. The pupils' and the parents' needs do not seem to have been met. The document may not only be too complex for the pupils, but may also be incomplete, because certain kinds of information that would be useful to them may not be included. The language of the document is the language of the instruction, but it is unlikely that the content will be readily comprehensible for the average 13-year-old, who should be reading the document a year before the actual test.

Norway: One set of test specifications exists; these are mainly intended for teachers, test developers and others who need information about the test. Since the specifications are available to the public, anyone interested may access them. Specific test-taker specifications have not been developed. However, teachers are instructed to inform the pupils about the test, and to make sure

they have been introduced to the sample tasks and the previous year's test before they take the national test of English.

Is there a description of the test taker?

Slovenia: There is no explicit description of the test takers. However, there are other documents describing the Slovenian educational system, school curricula and such, which provide a detailed description of the test takers.

Norway: The national test has been developed for all Year 5 students attending school in Norway. No further description of the test taker exists.

Are the constructs that are intended to underlie the test/subtest(s) specified?

Slovenia: The construct that the test is intended to assess is based on a functionalist view of language, language use and language proficiency, and is closely related to the theoretical framework of foreign language competence described in the CEFR.[5] Such a view relates language to the contexts in which it is used and the communicative functions it performs. In the case of this test, the ability to communicate includes, for example, the ability to comprehend texts, to interact in writing and to express one's ideas. The test assesses skills in reading and listening comprehension, written production and language use. The oral skills of the pupils are not tested.

Norway: The construct upon which the national test of English for Year 5 students is based is specified. The test assesses reading (understanding main points and details), vocabulary (common words in a context), and grammar (choosing the correct grammatical structure in a context; for example, singular/plural form of nouns, present form of verbs, personal pronouns).

Are test methods/tasks described and exemplified?

Slovenia: The test methods are described in the test specifications and exemplified through a number of test tasks available online. There are a variety of selected-response item types (e.g. multiple-choice, banked and unbanked gap-fill, matching and transformation) for assessing reading and listening skills, and language use; and open constructed-response items for assessing writing skills.

Norway: The test and the item formats are described in the guidelines for teachers. Sample tasks and the previous year's national test are available online (www.udir.no/vurdering/nasjonale-prover/engelsk/engelsk/). Teachers are encouraged to let their students do the sample tasks before they take the test in order to ensure that they know how to respond to the various item formats.

5 The Common European Framework of Reference for Languages: Learning, Teaching, Assessment: http://www.coe.int/t/dg4/linguistic/cadre_en.asp

Is the range of student performances described and exemplified?

Slovenia: In order to clarify the scoring criteria for the subjective marking of written texts, teachers/raters are provided with examples of a range of pupils' written performances at annual standardisation sessions. The Chief Examiner, assisted by colleagues from the National Testing Team for English, sets the standards for the marking; these are passed on to teachers/raters, who then mark the written scripts produced by the pupils at the schools. The National Testing Centre receives 10% of the pupils' written scripts, from which the Chief Examiner and her colleagues from the National Testing Team for English select the scripts that represent *excellent, adequate*, *average*, and *inadequate* performances. Next, the selected scripts are graded, discussed and compared by all of the members of the National Testing Team for English. Finally, a consensus mark is reached for each script. Once the team has reached an agreement, they record the reasons for each of their decisions, usually by writing justifications for each grade and allocating a certain number of points for each criterion/descriptor. The standardisation sessions, which are usually held a month before the test, take place every year in locations across the country in order to reach as many teachers as possible. It is strongly recommended that both novice and experienced teachers/raters attend these meetings.

Norway: Since the national test in English for Year 5 students does not test productive skills (speaking and writing), there are no examples of pupil scripts signifying different levels of achievement. However, the range of pupil performance in terms of total score is described. The Norwegian Directorate of Education and Training have asked the test developers to construct a test that discriminates between pupils at all levels. This means that the final distribution of test scores is expected to follow a normal curve, with the average pupil answering approximately half of the items correctly. It is important to explain this fact thoroughly for head teachers, teachers and parents, since most pupils do well on school tests. This is spelt out in many documents; for instance, it is specified in the teachers' guidelines that pupils who answer 50–60% per cent of the items correctly have done a good job. Less than 0.2% of the pupils obtain the maximum score on the test.

Are marking schemes/rating criteria described?

Slovenia: The marking scheme for each live test is available after the test has been administered. The marking scheme includes all the answers, including tapescripts and the writing rating scale. The benchmark scripts used in the standardisation meetings are not available to all stakeholders, only to the teachers who attended the annual standardisation session.

Norway: Norwegian pupils take the national test of English online. The items are scored correct or incorrect automatically. This means that no marking schemes or rating criteria for teachers exist.

Is the test level specified in CEFR terms? What evidence is provided to support this claim?

Slovenia: The National English Language Curriculum states that Year 6 pupils should achieve level A1. In 2008, the National Testing Centre started a project with the aim of aligning all national English language examinations to the CEFR; however, this process was only partially completed by August 2012. The project included 11 language experts and an international consultant. In defining cut scores, either the Angoff or Basket method was used, or a combination of the two. The project strictly followed the good practice principles for aligning tests to the CEFR, as defined in the Manual for Relating Language Examinations to the Common European Framework of Reference for languages: learning, teaching, assessment (Council of Europe, 2003). In the course of empirical validation, the test has been subjected to various classical and IRT-based procedures for the purpose of internal validation. To date, cut scores for reading, listening, and language use have been established for A1 and A2. As the project is not yet finished, no cut scores are currently available, and the CEFR statements have not been used in the reporting schemes to pupils. However, the curricula for the English language include language standards that have been aligned to the CEFR reference levels (Pižorn, 2009).

Norway: When the test was first developed in 2003/2004, the test developers linked it to the CEFR. This was done by basing test items on curriculum goals, as well as CEFR statements for the relevant skills. Most items were developed to measure A2 competence, while a few items were developed to mirror competence at A1 and B1 levels. In 2004, a major standard setting project was undertaken. A test-centred method, the Kaftandjieva and Takala compound cumulative method (which is a modification of the well-known Angoff method) was used. The project involved 20 judges assessing the CEFR level of several hundred test items. Cut scores were established for A1, A1/A2, A2, A2/B1 and B1. In 2004 and 2005, pupils' results were given in the form of a CEFR level or an in-between level, but some stakeholders considered it too complicated to have different scales for the different national tests. The Norwegian Directorate of Education and Training decided, therefore, that all national tests had to report test results in points, one point representing one correct answer. This means that no cut scores have been established. In addition to curriculum goals, the CEFR statements are still used as a basis for test and item development, but no standard setting procedures for this test have been applied since 2004.

Test Design and Item Writing

Do test developers and item writers have relevant teaching experience at the level the assessment is aimed at?

Slovenia: Test developers and item writers include a group of language teachers working at various primary schools across the country, a counsellor for English from the National Education Institute and an English language expert from the university. This team was put together by the National Testing Centre and the Ministry of Education, which made an effort to select highly motivated teachers who had a number of years of teaching experience and, ideally, had been trained in language testing. These positions are for four years, but the decision makers try to keep a certain number of senior members on the team while recruiting new ones to ensure that what has been learned from experience is not lost.

Norway: The team developing the national test of English includes professional test developers, a teacher, teacher trainers and an artist who draws pictures for the tests. All of the test developers except one have a background as English teachers, and together they cover primary, lower-secondary and upper-secondary school. The primary school teacher on the team works 40% of the time on the national tests and 60% at a local primary school teaching English.

What training do test developers and item writers have?

Slovenia: The test developers and item writers working on the test generally have considerable teaching experience but vary considerably with regard to their training in language testing. It is therefore recommended that they obtain extra training, either in the country or abroad at well-known institutions specialising in language testing. Thus far, several members of the National Primary School Testing Team for English have been trained at one of the best UK universities for language testing.

Norway: All of the test developers and item writers except one are English teachers, most with degrees from a teacher trainers' college or university. Some of the test developers also have a background in theoretical studies in second language learning. Most of the test developers have attended international courses focussing on language testing and item writing, and those who have not attended such a course will do so soon. More importantly, the test developers work as a team. Individual test developers make suggestions for items, the items are scrutinized by the team, and changes are suggested and discussed.

Are there guidelines for test design and item writing?

Slovenia: When Slovenia started to design national foreign language tests for primary schools, there was little or no language assessment expertise available. It was decided, therefore, that a document that would give an overview of language assessment for this age group should be published. In 2003, the National Primary School Testing Team for English worked with an international language testing expert to produce a book addressing most of the testing issues; it was designed with the intention of providing general guidelines for novice test designers, who were usually language teachers with very little knowledge and experience in language testing.

Norway: No self-made written guidelines for test design and item writing exist, other than what is included in the test specifications with regard to test content and item formats. New employees have been taught by their more experienced colleagues and also attended international courses in item writing and language testing. The team has access to international literature focusing on item writing. The plan was to sum up the developing team's experiences and write specific guidelines for item writing before the 10th anniversary of the national tests of English in 2012.

Are there systematic procedures for review, revision and editing of items and tasks to ensure that they match the test specifications and comply with item writer guidelines?

Slovenia: The National Primary School Testing Team for English designs a detailed work plan and each member has to design a certain number of tasks, which are first checked to determine whether they conform to the criteria in the Test Structure document. After the round table discussion at which test items are either kept or discarded, special attention is paid to the content quality, clarity, sensitivity and bias of proposed test items. The tasks that have been accepted are then revised and edited, and prepared for pretesting. After trialling the selected tasks on approximately 100 pupils, the two traditional measures, i.e. the facility value and the discrimination index, are calculated. Writing tasks are pretested on a smaller number of students with a wide range of language levels in order to ensure that the sample of language produced contains most of the features required by the task.

The item bank tasks are organised according to a number of variables (language skills, topics, text source, length, task type etc.), which is helpful in structuring the whole paper and in this way complying with the test specifications.

Before the test is finalised, all the items associated with each and every

task are reviewed by a native speaker, two practising English language teachers who teach the test age group and a testing expert, in this case, the National Testing Centre coordinator for languages.

Norway: The test specifications contain an overview of the content of the test, describing approximately how many items measuring each of the skills and how many representing each item format are to be included in the test. All items are tagged for item format and for the skill being tested. When pilot versions of tests and real tests are constructed, test developers ensure that this model is followed. External reviewers comment on suggestions for all pilot versions of the pilot tests and real tests. Such comments are always discussed and considered. When final tests are agreed upon, a period of extensive checking takes place. A final proof reading is done. Pictures, layout, answer keys and automatic scoring are checked independently by at least two persons. Last, but not least, the guideline for teachers is checked and revised every year, both by persons in the directorate and by test developers.

What feedback do item writers receive on their work?

Slovenia: The National Primary School Testing Team for English are the item writers and the editing team, which means that each test task and each test item is reviewed by six people. It is not surprising, therefore, that the only information that they receive from outside their group is the statistical data from the pre-testing and the data from the short questionnaires that are attached to each of the test tasks. However, some valuable feedback may also come from the test expert and the two practising teachers who review the whole test/s.

Norway: Since item writers work in teams, they have extensive feedback from colleagues when developing items. The team also have meetings to consider the piloting and test data. In these meetings, there are discussions about, for example, why certain items have positive / less positive discrimination indices, and why some items are difficult/easy.

Quality Control and Test Analyses

Are the tests pilot tested? What is the normal size of the pilot sample, and how does it compare with the test population?

Slovenia: Individual test items are pre-tested on 100 students of similar age; however, there is no large-scale field test due to a lack of resources. The normal size of the pre-test sample is around 100 Year 6 students. These students are of similar age and language competence as the live test population. The test population is approximately 13,000 to 14,000 students.

Norway: The tests and items are piloted on approximately 3000 Year 5 students, one year before test administration. This means that the pupils taking part in piloting will not encounter the same items in the real test. Statistics Norway (www.ssb.no) selects the schools that are to take part in the piloting in order to ensure a sample of pupils that reflects the entire population of pupils. For example, pupils from urban and rural areas are included, as well as pupils from small and large schools.

How are changes to the test agreed upon after the analyses of the evidence collected in the pilot?

Slovenia: Generally, all items with a facility value of above 85% and less than 20% are discarded, and the same is true for test items whose discrimination index is below +.4. Items testing writing skills, such as letters, descriptions and postcards are pre-tested on a smaller number of pupils, but with a wide range of language levels in order to ensure that the sample of language produced contains most of the features required by the task.

Norway: Several different test versions are piloted, and the final test will contain items from most versions. The final test will include all of the agreed upon item formats, and items measuring all of the skills the test is intend to measure; the number of items testing each skill is specified. Items for inclusion in the final test are selected and agreed upon by the test developers and the statistician who has performed the IRT analysis.

If there are different versions of the test (e.g., year by year), how is the equivalence verified?

Slovenia: Every year, two parallel versions of the test have to be developed. Round-table discussion of the test items, the pre-testing of individual test tasks, content analysis and other procedures support the construction of parallel tests. Due to the small samples used in the pre-testing of tasks/items, it is not possible to reliably predict item behaviour in the live tests.

Norway: From 2004 to 2011, three parallel versions of the national test of English for Year 5 students were developed every year. IRT analysis and careful selection of items made it possible to construct three parallel tests. The Ministry of Education decided in 2011 to initiate research into the changes in the competence of the population over time. The test developers have been instructed, therefore, to develop only one test version (instead of three), starting in 2012. In addition, a set of anchor items is to be selected. This means that two test versions are now required: one main version, which will be administered to 99% of the students; and one anchor version, in which some of the items in the main

test have been replaced by the anchor items. The items selected as anchor items have been piloted together with the other items, and have behaved in more or less the same way as the items that they replace in the main test.

What statistical analyses are used?

Slovenia: Both Classical Test Theory (CTT) and Item Response Theory (IRT) are employed to analyze test item data. CTT analyses provide p-values, item-total correlation, maximum scores, mean scores, point-biserial statistics and multiple-choice option statistics. IRT analyses provide item parameter estimates, fit statistics, ability and item difficulty estimates, and differential item functioning statistics. In addition, a variety of other statistical analyses, including cluster analysis, factor analysis and multiple regression, were used to help understand the underlying constructs measured, as well as the pupils' performance profiles.

Norway: Both Classical Test Theory analysis (CTT) and Item Response Theory analysis (IRT) are used in the data analysis. Both types of statistics provide information such as the p-values, discrimination indices and reliability measures. IRT analysis makes it possible to analyse items from the various pilot version on the same scale, as well as to construct parallel test versions. In addition, IRT analysis provides fit statistics and information about test taker ability and item facility on the same scale.

What processes are in place for test takers to make complaints or seek reassessments?

Slovenia: If a test taker is unhappy with his/her test score, he/she can request a rescore. Full details regarding how to proceed with the rescore application are provided on the National Testing Centre website. If a test taker believes an error has been made in any part of the test that may have affected his/her score, their teacher can complete the Item Challenge Form.

Norway: Teachers have access to their pupils' test answers and test scores. Teachers log on to the Directorate website and view online versions of the tests their pupils have completed. Complaints are addressed to the Norwegian Directorate of Education and Training. As the scoring procedure is automatic, there are no complaints in connection with scoring. This means that reassessments do not happen. What may happen, however, is that teachers question specific items or item formats.

Conclusion

In the field of foreign language learning and teaching for young learners, research related to assessment has been generally neglected. Few studies have been dedicated to large-scale assessment processes, and hardly any to test validation procedures related to large-scale national foreign language tests to young learners. This validation study of two national tests of young English language learners has been an attempt to validate the two national tests in Norway and Slovenia by applying the EALTA guidelines principles.

Below, we will first identify and discuss the areas in which the development and application of these national tests has been undertaken in accordance with internationally recognized standards of good practice in language testing and assessment, as defined by EALTA guidelines for good practice. Second, we will focus on the issues that have been identified as needing improvement and finally, we will provide some recommendations of how to improve test development and application procedures. However, before the final discussion we would like to acknowledge the role of the EALTA guidelines for good practice in making it possible to validate the test and offer recommendations for improvement.

The tests' purpose(s) seem to be clearly and transparently defined and are available online to all stakeholders. Both national tests have the same core goal of providing additional information about students' achievements in English. All stakeholders in both assessment contexts have been identified and their responsibilities acknowledged. Test providers in both countries put considerable effort into preventing the misuse of the tests and the misinterpretation of the results by enrolling only pupils of the same year, providing detailed information about the test, and giving thorough feedback to pupils and teachers.

Test specifications have been developed in both countries, which are now available in the form of one or more online documents. These seem to be comprehensive and helpful documents for test designers and other stakeholders involved in the assessment process. They have gone through multiple revisions in response to feedback from various sources, resulting in annual updates and revisions.

The test constructs for both tests are based on the view that relates language to the contexts in which it is used and the communicative functions it performs. Thus, the focus of the test tasks is on the assessment of pupils' communicative language competence.

Full and accurate descriptions of test methods and tasks are provided for both national tests and relevant sample tasks are available online. Teachers

are also encouraged to let their pupils do the sample tasks before the live test, which should reduce the influence of test methods on students' scores (Alderson, 1995, p. 44).

It seems that test design, item writing and quality control procedures follow the established high standards in language testing and assessment. The testing teams involve testing experts and practising teachers who can relate to the test takers' needs and interests. Item writers/test designers work in teams that discuss all items/tasks in detail before pre-testing.

A number of statistical analyses are performed on data from both national tests and made publicly available to all stakeholders or anyone interested in the results of the national tests.

The areas that need to be improved and further researched refer to the publication of school results in the media and the high expectations of head teachers and parents who put extra pressure on teachers and pupils, and on transforming low-stakes tests into high-stakes tests. In Slovenia, parents may need more information regarding the test itself and helpful guidelines for supporting their children's foreign language learning. There is also no evidence as to what extent stakeholders have been informed about the test, or as to whether this information has satisfied different stakeholders' needs.

Today's goal of language learning and teaching is to develop students' language competence, involving listening, speaking, reading and writing skills. Neither of the national tests studied assess pupils in all four skills; therefore, they do not provide a whole picture of pupils' language competence. Although it is made clear that teachers need to assess the missing skills formatively, it is, however, a fact that in the long run what is tested is usually what is taught, so the test providers should be aware of this potential "unintentional" washback effect. Avoiding assessing oral skills is common practice, unfortunately, in external testing (McKay, 2006, pp. 176–177), in spite of the fact that oral language constitutes a central core of young learners' curriculum and instruction time. McKay (ibid.) warns that failure to assess oral language and assessment of language learning through reading and writing denies the essence of young learners' language learning. What is even more worrying is the fact that policy makers tend to rely much more heavily on summative test results than on teachers' formative grades. Furthermore, the test report that pupils receive in Slovenia does not include any detailed information of the test structure or language skills assessed, only the aggregated score in points. Thus, pupils who are good speakers but less proficient readers or writers, or who are less competent in using grammatical structures accurately, will not be able to show their mastery in this important language skill.

The range of written performances by pupils is available only in Slovenia as the test in Norway does not include any productive skills. The Slovenian benchmark scripts are available to all teachers who attend the annual standardisation meetings, and recently have a few sample scripts representing specific rating-scale descriptors been included in the moderated answer key, which is available to all teachers across Slovenia. However, attending standardisation meetings is voluntary and left to the teachers, which may influence the inter- and intra-rater reliability results.

Both national tests have been aligned to the CEFR following the strict procedures defined in *The Manual for Relating Language Examinations to the Common European Framework of Reference for Languages* (Council of Europe, 2003).Unfortunately, this was a one-off event and the cut scores are not established for the current tests; nor have pupils' results been reported using the CEFR reference levels.

In Norway, pretesting involves a large sample of students, while this is not the case in Slovenia, where decision makers should make it possible for the testing team to pilot test tasks on a representative sample of students, since this may be one of the reasons that the test index of difficulty varies from year to year. Another unresolved issue is the fact that, due to the online publications of all live papers, no anchor items can be included in the test as fixed parameters, making it impossible to monitor pupils' achievements from a longitudinal perspective.

This validation study showed that there are a number of issues in the development and administration and use of the national tests that should be researched and ultimately improved. We recommend that the test designers in Slovenia find the most appropriate way of piloting test items and test papers on a larger sample of the target population. This can either be achieved by establishing a piloting system within The National Testing Centre or by changing the paper-based format of the test to a computerized one.

Another recommendation is that all stakeholders should be aware that publicising test results openly does not serve the students' needs. We believe that tests need to be yearly aligned to the CEFR and should move away from norm-referencing the students. If criterion-referenced tests become accepted by the stakeholders involved in language learning and teaching processes, students will learn to compete with the criteria (the CEFR levels) and there will be less competition among students and/or parents.

Both national tests should develop national speaking tasks that could be available to teachers with benchmarks and appropriate feedback for the students and their parents. In Norway, carefully developed writing tasks and

benchmarks would be useful for teachers, students and parents. Standardisation meetings should be organized and made compulsory for all teachers. The national test of English for Year 5 students in Norway should assess students' listening skills, and cover more of the curriculum goals.

References

Alderson, J. C., & Banerjee, J. (2008). *EALTA's Guidelines for Good Practice: A test of implementation.* Paper presented at the 5th Annual Conference of the European Association for Language Testing and Assessment, Athens, Greece, May. Retreived from http://www.ealta.eu.org/

Alderson, J. C., & Wall, D. (1993). Does Washback Exist? *Applied Linguistics, 14*(2), 115–129.

Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing, 27*, 51–72.

Alderson, J. C., & Banerjee, J. V. (2008). EALTA's guidelines for good practice: A test of implementation. Paper presented at the 5th Annual Conference of the European Association for Language Testing and Assessment, Athens, Greece, May. Retrieved 24.12.2008 from http://www.ealta.eu.org/conference/2008/programme.htm

Breen, M. P., Barratt-Pugh, C., Derewianka, B., House, H., Hudson, C., Lumley, T., & Rohl, M. (Eds.) (1997). *Profiling ESL Children: How Teachers Interpret and Use National and State Assessment Frameworks*, Vol. I. Canberra: Department of Employment, Education, Training and Youth Affairs.

Brumen, M., Cagran, B., & Rixon, S. (2009). Comparative assessment of young learners' foreign language competence in three Eastern European countries. [Article]. *Educational Studies (03055698), 35*(3), 269–295. doi: 10.1080/03055690802648531

Council of Europe (2004). *Reference supplement to the preliminary version of the manual for relating examinations to the Common European Framework of reference for Languages: learning, teaching, assessment. DGIV/EDU/LANG*. Strasbourg: Language Policy Division.

Council of Europe (2001). *Common European Framework of Reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe (2003). *Relating Language Examinations to the Common European Framework of Reference for languages: learning, teaching, assessment (CEF).* Manual: Preliminary Pilot Version, DGIV/EDU/LANG. Strasbourg: Language Policy Division.

De Jong, J., & Zheng, Y. (2011). Research Note: Applying EALTA Guidelines: A Practical case study on Pearson Test of English Academic. Retrieved 30.6.2012 from http://www.pearsonpte.com/research/Documents/RN_ApplyingEALTAGuidelines_2010.pdf

Edelenbos, P., & Johnstone, R. (Eds.) (1996). *Researching Languages at Primary School*. London: CILT.

Educational Testing Service. (2002). *ETS standards for quality and fairness.* Princeton, NJ: ETS.

European Association for Language Testing and Assessment (2006). Guidelines for Good Practice in Language Testing and Assessment. Retrieved 26.6.2012 from http://www.ealta.eu.org/guidelines.htm

European Commission. (2008). Key Data on Teaching Languages at School in Europe. (pp. 136).

Retrieved from http://eacea.ec.europa.eu/education/eurydice/documents/key_data_series/095EN.pdf doi:10.2797/12061

Eurydice, E. P. (2009). *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results*. Brussels: Education, Audiovisual and Culture Executive Agency.

Graddol, D. (2006). English Next. *Why global English may mean the end of 'English as a Foreign Language'* (pp. 132). Retrieved from http://www.britishcouncil.org/learning-research-english-next.pdf

Hambleton, R. (2010). *A new challenge: Making test score reports more understandable and useful*. Paper presented at the 7th Conference of the International Test Commission. Hong Kong.

Hughes, A. (1989). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Interpreting the PTE Academic Score Report. Retrieved from http://pearsonpte.com/

Leung, C., & Teasdale, T. (1997). What do teachers mean by speaking and listening: A contextualised study of assessment in the English National Curriculum. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Louma (Eds.), *New contexts, goals and alternatives in language assessment* (pp. 291–324). Jyväskylä: University of Jyväskylä.

Low, L., Brown, S., Johnstone, R., & Pirrie, A. (1995). *Foreign Languages in Primary Schools.* Stirling: Scottish Centre for Information on Language Teaching and Research, University of Stirling.

Mckay, P., Hudson, C., & Sapuppo, M. (1994). NLLIA ESL Bandscales. In P. McKay (Ed.), *NLLIA ESL Development: Language and Literacy in Schools*, Vol. I (pp. B1-D52). Canberra: National Languages and Literacy Institute of Australia.

McKay, P. (2000). On ESL standards for school-age learners. *Language Testing, 17*(2), 185–214.

McKay, P. (2006). *Assessing Young Language Learners*. Cambridge: Cambridge University Press.

Pižorn, K. (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*, (pp. 87-101). Arnhem: Cito, Institute for Educational Measurement: Council of Europe: European Association for Language Testing and Assessment, EALTA.

Rea-Dickins, P. (2000). Assessment in early years language learning contexts. *Language Testing, 17*(2), 115-122. doi: 10.1177/026553220001700201

Biographical note

Karmen Pižorn is Assistant Professor of English Language Teaching Methodology at University of Ljubljana, Faculty of Education, Slovenia. She holds a Ph.D. in English language teaching methodology (University of Ljubljana, 2003). She has taught various courses at BA, MA and PhD levels related to English language teaching methodology, language assessment and general English. Her research interests include all aspects of language testing, teaching foreign languages to young learners, and other issues involved in learning and teaching foreign languages in general.

Eli Moe is working at the University of Bergen, Norway. She is leading several test development projects: Digital national tests in English for Norwegian 5th and 8th graders (commissioned by the Norwegian Directorate of Education and Training, an agency of the Norwegian Ministry of Education and Research); Diagnostic tests in English for Norwegian 3rd and 11th graders (commissioned by the same agency); Digital tests in Norwegian as a second language for adult immigrants (commissioned by Vox, Norwegian agency for Lifelong Learning). Eli Moe is also leading an ECML project focussing on young learners and the language of schooling. She has been involved in the translation of the Common European Framework of Reference to Norwegian, and been a part of the team developing the Norwegian version of Dialang.