Author(s): Pižorn, Karmen; Huhta, Ari

Title: Assessment in educational settings

Year: 2016

Version:

Please cite the original version:

Pižorn, K., & Huhta, A. (2016). Assessment in educational settings. In D. Tsagari, & J. Banerjee (Eds.), Handbook of Second Language Assessment (pp. 239-254). De Gruyter. Handbooks of Applied Linguistics, 12. https://doi.org/10.1515/9781614513827-017

Assessment in educational settings

Karmen Pižorn and Ari Huhta

1. Introduction

In this chapter we focus on large-scale national foreign language assessments (LS-NFLAs) that are related in some way to a country's educational system. In the first part of the chapter we define what we mean by LS-NFLAs and discuss features that characterize them. We then give an account of the changes in these assessments in the past few decades and discuss some current trends in them. In the second part, we focus on their uses and purposes, their development phases and administration issues, their common structure, the sustainability of good testing practice, the influence of politics and the washback effect of the assessments on teaching and learning. Finally, in the third part, we discuss how national assessments are likely to evolve in the future.

We begin by defining what kind of large-scale assessments are covered in this chapter. We focus on assessments that are related to the educational system of a fairly large entity such as an entire country or a major part of a country, for example, a state or a region,. Because of the wide geographical coverage of these assessments and also because they usually have an official status, they are often called national assessments. National assessments are typically centralized, standardized and based on a national or regional curriculum, that is, they are designed outside the school by an assessment organization or an examination board, although some locally designed assessment instruments may also be used as part of the national assessment. These assessments are almost always carried out by using tests, although information from other types of assessments such as the portfolio or assessments by the teachers may complement the test scores (therefore, we often refer to national assessments also as national tests or national examinations below). National assessments often take place at the end of a major educational level such as the end of primary, lower secondary or upper secondary education, although in some countries they may be administered more often.

It is important to consider the purposes of national assessments, as they can differ considerably and also their role and impact on the educational system can differ depending on the use of the assessment results. The uses of national assessments fall into two main categories: assessment of the learners and assessment of the educational programmes (see Brown 2013). The use of national assessments as a way to measure how much of the curriculum the students have achieved by the end of some important unit of studies such as compulsory education is probably the most common use of national assessment / examination results. That is, the students' learning is evaluated and they are given grades that indicate how well they have achieved the goals of whatever education they have just completed. These tests are typically quite high-stakes for the students, since the results are often used for gaining entry into the next level of education. For example, the final primary school examination may determine the kind of secondary education the student can enter (e.g. vocational or academic). It should be added that the results of school-leaving examinations can also be used by educational authorities for the other major purpose of assessment, which is discussed below: to evaluate the quality of an educational programme (e.g., a national curriculum).

There is, however, another type of national assessment that is not concerned with awarding individual students grades. The sole purpose of these assessments is to investigate the quality of an educational programme or system so that its current state could be determined and also to find out how much variation there is in the performance of different schools, regions or genders, and whether certain aspects of the curriculum are mastered better than others. International examples of this type of assessment are the PISA studies (however, unlike national assessments with a similar purpose, the PISA studies are not based on any national curricula). Based on the findings, the authorities may decide to introduce changes in the educational system, for example, a revision of the curriculum or allocation of more resources to schools with lower results. It is important to understand that this type of national assessment is closer to a research study than to a national school-leaving examination. Therefore, the administration of a national assessment intended purely for programme evaluation purposes can differ radically from a national achievement test. Because one (perhaps the main) purpose of a national *achievement test* is to award grades to individual students, it is imperative that the test is administered to all eligible students. To do otherwise would be unfair to the students, especially if the test results are important for their further education or entry into labour market. The fact that the results from such national achievement tests are in some countries also used for informing the educational authorities (and possibly also the schools) can be seen as only a secondary use of the test results.

When a national test is not used for grade-giving and other such high-stakes purposes for individual students but merely to gather data to assess the quality of an educational programme, it is not necessary to administer it to all students. It is enough to take a statistically representative sample of schools and students, which is less expensive and which yields basically the same information about the educational system as testing all schools and all students would. Unlike the achievement tests, these assessments are not high-stakes for the students, although the results can be very important for politicians, educational decision makers and curriculum planners.

In our review, we focus on large-scale national foreign language (FL) assessments whose principal purpose is to measure students' achievement in primary and secondary education, and to give them grades based on their test results. These tests are often school-leaving examinations and may also be used for assessing the quality of a country's educational system. This review predominantly focuses on the national FL assessments in Europe but the examples, though contextualised in Europe, reveal themes that are relevant globally. We also refer to selected examples from other regions of the world.

2. Historical development of national foreign language assessments

Next we briefly discuss selected major trends discernible in the development of large-scale national FL assessments in the past few decades: communicativeness, the CEFR, standardization, and professionalization.

It is obvious that national language assessments have become more communicative in the past forty years, as have language tests more generally (see, e.g., Kunnan 2008; Davies 2014). Most tests regardless of their purpose tended to measure knowledge about language till the 1970s because of the focus on errors and discrete elements of language, especially grammatical structures and vocabulary. The widespread use of multiple-choice and translation as test methods contributed to this emphasis. The testing of the productive skills

was not common in national assessments, partly for practical reasons and partly for a concern about their reliability.

Largely because of changes in how language proficiency was conceptualized, language testing started to change. Since the 1970s, language proficiency has no longer been viewed as knowledge about distinct aspects of language but rather as an ability to use language for various communicative purposes, in different contexts and with different interlocutors or recipients. Language tests nowadays attempt to incorporate features of real-life communication as much as is feasible.

The Common European Framework of Reference for languages (Council of Europe 2001) started as a purely European initiative, as its name suggests, but in the past decade it has become influential in other regions of the world, too, particularly in Asia, Middle East and South America. This document defines in quite some detail what it means to communicate in a FL and how ability to use language can be defined as levels of ascending skill and sophistication. In this way, the CEFR has supported and further contributed to the prominence of communicative language teaching and testing.

The CEFR has become increasingly important in high-stakes language testing. This is particularly evident in international language testing as even tests not based in Europe have considered it important to align themselves with the CEFR, presumably to ensure their relevance not only for their European users but also for other contexts in which the CEFR has become popular. For example, the US based TOEFL iBT has carried out such alignment procedures (Tannenbaum and Wylie 2008). The impact of the CEFR on language education and on language testing in Europe has been significant (Martyniuk 2011; Figueras 2012). Martyniuk and Noijons summarise their survey on the use of the CEFR across Europe: “The overall impression is that the majority of countries have already been trying to implement the CEFR for some time in the development of tests and examinations either for primary and secondary schools or for adult education” (2007: 7).

Language testing has become more professional in many countries during the past few decades. Interestingly, Figueras (2007, 2012) argues that the CEFR has played a part in this, for example, because of the need to link various examinations and assessments with its levels. The increasing standardization of national high-stakes examinations that will be described later in this chapter is one indication of that professionalization. Another sign is the creation of international organisations for professional language testers as well as for researchers, teachers and other people interested in language assessment. These include ILTA (International Language Testing Association), ALTE (Association for Language Testers in Europe) and EALTA (European Association for Language Testing and Assessment). These associations organize conferences and workshops that serve as arenas for disseminating research results and for exchanging information about good practice in language assessment. They also promote and sometimes even fund local activities such as training events that aim at increasing different stakeholders’ awareness of language testing issues and principles of good practice. An important aspect of the associations’ work is the design of codes of ethics and guidelines for practice that guide good professional conduct in language testing.

3. The Key Issues of large-scale foreign language national assessments

3.1 Test use and purposes of large-scale foreign language national assessments

According to Davies (1990) defining the test purpose is the first step in the test design process and must be asked and answered before we can decide upon the test content and test methods. This is because the purpose for which the test will be used influences what is tested and how it will be tested, i.e., which language skills, which topics and test methods, more or less specific language use etc. So, for example, a school-leaving examination in English as a FL will be indicating progress according to the objectives set by the national curricula, but may also be used for certification or form part of a process of program evaluation. It can even have a diagnostic function, which in turn may assist in selection decisions, for example, functioning as a university entrance examination. This is exactly the case with many European secondary-school leaving examinations (Eckes et al. 2005) where national examinations have different purposes. This can also be observed from the Eurydice[1] report (2009) where test purposes of national assessments are often varied and rarely one-dimensional.

In Europe, national assessment of students is becoming increasingly important as a means of measuring and monitoring the quality of education, and structuring European education systems (Eurydice 2009). According to Eurydice (2009), the national FL assessments are developed and influenced by national policy frameworks and contexts, and are usually part of assessment of other school subjects or areas of study. They should contribute to a more comprehensive picture of student knowledge and skills by providing additional information to parents, teachers, schools and the entire educational system. The national assessment systems within compulsory education levels have been introduced in almost all European countries over the last three decades, and have grown to an important instrument in the organisation of educational systems (Eurydice 2009). In fact, more than one third of European countries administer national assessments in foreign languages already at primary level, and 60% of countries assess FL proficiency of their students at the end of compulsory education (Eurydice 2009).

The analysis of the objectives and uses of national assessments including assessing foreign languages in the 35 countries/country communities at ISECD levels 1 and 2[2] reveals that the main objectives of such assessments are monitoring and evaluating educational systems (17 countries or over 40%) and examining whether the objectives set by the national curricula have been achieved (17 countries). These are followed by providing schools with the information on their students' achievements and offering a tool for their self-evaluation (seven countries), and informing all stakeholders about students' attainments (six countries). Surprisingly, providing teachers as one of the main educational stakeholders with extra information on their students' achievements is explicitly stated as an objective of national assessments only in five countries/country communities. Another interesting finding is that the idea *assessment for learning* (Black and Wiliam 1998) had quite a long history now; however, only four of countries/country communities clearly stated that providing learning opportunities for schools belongs to the main objectives of the national assessment in their educational context. There are also only four countries whose main objectives of the national assessment incorporate providing certification of attainment or making necessary policy changes. Only two countries reported utilizing national assessment for guiding the streaming of pupils, or providing diagnostic information on students' achievement, or informing parents of the pupil's summative achievement. It is interesting to observe that, with the exception of Finland, national assessments do not focus on monitoring the implementation of equality and equity in education.

Whether and to what extent the LS-NFLAs mentioned above actually measure the

stated purposes should be carefully investigated by taking into account the effect the test is intended to have in the real world. In other words, do tests test what they claim to test? To answer this we need to do a validation analysis of a particular test. Validation is concerned with the gathering of as much evidence as possible, which would support or refute the inferences that are made about test takers based on their performance on the test, and the decisions that are made about learners based on their test scores. According to Fulcher and Davidson (2009) the intended score meaning should be explicitly and carefully linked to test design, as otherwise it is not possible to relate the users' interpretation of the score to the decisions that they take on the basis of the score. Another issue that may undermine the appropriateness of tests is the lack of need for justifying the validity of a test among decision makers who are predominantly not language assessment experts. They are usually also not aware that it is not only high-stakes tests which require rigorous validation but also the low-stakes ones, if they are used nationally. Further, the difference between low- and high-stakes national FL tests cannot be universally and simply defined and national assessments may have unintended consequences from the original set objectives. Pizorn and Moe (2012: 81) report that the national FL assessments for young learners in two European countries are not supposed to directly influence the students' final grades or have implications on their future career. That said, the language teachers in these countries have expressed considerable concern about the pressure they are under from the head teachers and parents. This is especially so because despite opposition from the Ministry of Education, the test results are published in national newspapers along with the ranking of schools. It is, therefore, vital that more validation studies of national FL tests are performed and that these studies are open to international language testing communities' scrutiny. Chapelle, Jamieson and Hegelheimer (2003: 413) propose doing regular analyses, which would make explicit the links between the components of test purpose (the inferred test use and its intended impact) and the design and validation decisions.

Not many LS-NFLAs have been openly validated in peer-reviewed academic journals, so there is not much evidence whether these instruments test what they claim to test. When investigating different types of validity (content, face, construct etc.), it is content validity of the test that may have a strong influence on what is being taught in the classroom and may narrow down the curriculum goals and objectives. Content validity includes any validity strategies that focus on the content of the test. To demonstrate content validity, test designers should investigate the degree to which a test is a representative sample of the content of whatever objectives or specifications the test is originally designed to measure (Anderson 1975: 460; Hughes 2003).

3.2 Test development and test implementation

LS-NFLAs are gradually adopting standardization procedures including the use of quantitative and qualitative methodologies for item and test analysis, the detailed description of exam organization and testing conditions, as well as rater training and monitoring.

The whole process of test development consists of designing, planning, item-writing, pre-testing, editing and printing, distribution to schools, marking, setting pass marks if applicable, analyzing and reporting the results and the overall evaluation. Test development is a standardized procedure, which consists of various independent stages, which come at a specific time and place in the development process, yet each stage functions only in relation to the others. Tests should be developed according to clearly defined specifications. Language assessment requires measuring instruments constructed on the basis of sound

psychometric criteria and appropriately chosen test methods. Reforming language assessment practices needs to be embedded in continuing efforts of establishing the highest quality, encompassing a number of aspects ranging from the objective measurement of examinee proficiency and item difficulty to precise definitions of test administration conditions and scoring systems (Alderson 2004; Weir 2004). So, for example, Alderson and Pizorn (2004) point out that designing papers and detailed marking schemes should not be developed successively but simultaneously. If the stages of the test development, which are inextricably interwoven and dependent on the idiosyncrasies of an educational system, are not interrelated, the system will not enable alterations and improvements and cannot be transparent from the point of view of test developers, administrators and test users. Constraints that test developers may encounter at national levels are numerous, and range from time, human and financial resources, to political interference.

In the remainder of this section, we will discuss challenges that testing teams are faced with. It is not uncommon that testing team members, who frequently consist of ordinary language teachers and/or few (university) language specialists, have very little or no expertise in constructing FL assessments on a national level, which can only lead to inappropriately designed tests, if not to a total disaster. In the 90s many countries of East-European block started to set up national large-scale examinations and assessments and some test designers received a proper training in language assessment while others did not. In some cases language teachers had been trained as items writers for years, but were ultimately not selected to design the 'real' examination tasks and were replaced by total novices (Eckes et al. 2005). Test developers not only need the professional skills to produce good measurement instruments, but also to be able to apply these skills in creative ways, and to novel situations.

Another burning issue is the time one has to set up the assessment system. It takes time to build a national assessment system and if the tests are not well prepared and/or stakeholders are not well informed and far enough in advance, the consequences may be damaging. The decision-makers in two East-European countries were forced to postpone the introduction of a new examination due to the students' and/or teachers' protests (Eckes et al. 2005; Pižorn and Nagy 2009). As Buck (2009:174) observes there is often considerable pressure for test developers to complete a test as quickly as possible, simply to make the project go ahead and/or make them cheaper but with more or less unpredictable and fatal consequences.

Piloting is another issue that in the context of national assessments is treated in different ways. There are very few countries where all test items are piloted on a sufficient number of test-takers who have similar characteristics to the target test population. Cost is often cited as a barrier to piloting; it is an expensive activity. Yet a lack of funding is not the only barrier to piloting but also the assessment culture. Eckes et al. (2005) report on the experience of one country where no central piloting of examination items was planned, as some stakeholders feared that pretesting could jeopardize security. The result was that the items for the centrally designed papers at both levels were written behind desks, based upon the expertise of the individual item writers. Though Buck (2009: 174) is right in saying that there is no alternative to piloting tests as otherwise we have no idea whether the items are working properly, if the test providers on national level do not understand the need for piloting, language test designers have a very difficult job in persuading decision makers to offer support. Thus, many testing teams have to live with small-scale piloting or pilot test items in their free time, without the decision makers' support.

3.3 The structure of the large-scale national foreign language assessments

Currently, most of the national FL assessments show a number of advances in language assessment and have moved away from traditional knowledge-based tasks measuring rote learning. These can be seen through their considerations of (a) the theoretical view of language ability being multi-componential, (b) the correlation and the impact of test tasks and test taker characteristics on the test scores, (c) the application of sophisticated measurement instruments including more and more advanced statistical tools, and (d) the development of communicative language tests that incorporate principles of communicative language teaching (Bachman 1990; Bachman and Palmer 1996). Hence, students are assessed through more authentic tasks, which measure their reading and listening comprehension skills, writing and speaking interactional and transactional skills, and the use of vocabulary and grammar in context. While decades ago test takers tested on a national level were supposed to translate sentences, recite grammatical rules and dialogues by heart, they are now expected to be able to skim, scan and infer information from an authentic newspaper article, talk about their views and attitudes to the topics of their interests and relevance and write a letter of application or complaint, as well as compose a narrative or a discursive essay. How far these tasks are authentic and appropriate for the targeted audience of test takers has to be investigated in each individual educational context as each has its own idiosyncrasies. For example, Table 1 shows the structure of a secondary-school leaving examination in English at the basic level in Slovenia.

Table 1: The Structure of the Slovene Matura secondary-school leaving exam in English

Paper No.	**Skill/knowledge**	**Time**	**Weighting**	**Marking**
1A	Reading comprehension	35 min	20%	Centralised
1B	Knowledge and Use of language	25 min	15%	Centralised
2	Listening comprehension	Up to 20 min	15%	Centralised
3A	Writing (short, guided) (150-180 words)	30 min	10%	Centralised; double marking;
3B	Writing (essay) (220-250 words)	60 min	20%	Centralised; double marking;
4A	Speaking, picture discussion; the tasks prepared centrally;	Up to 20 min	20%	Internal by the teachers using centrally designed criteria
4B	Speaking Teacher-prepared guided task			
4C	Speaking; interpretation of a literary text and a follow-up discussion on the text topic; the tasks prepared centrally;			

The exam may be characterized as performance assessment reflecting the cognitive-constructivist view of learning. In performance assessment, real life or simulated assessment exercises are used to elicit original responses, which are directly observed and rated by a qualified judge. As such, performance-based tests could serve as driving-force for a thinking-oriented curriculum geared towards developing higher order thinking skills (Resnick and

Resnick 1992). However, investigations into the impact of tests upon teaching and learning (e.g. Alderson and Wall 1993 and Wall 2005) show that this is too simplistic a view.

3.4 Sustainability of good testing practice

From the 90s up to the middle of the 21st century, many countries of Central and Eastern Europe went through dramatic changes in language learning, teaching and assessment. In most instances (e.g. the Baltic states, Hungary, Slovenia), assessment reforms started as projects and were supported financially, and what is more important professionally (Eckes et al. 2005). The aims of these projects were similar: to develop a model for a transparent and coherent system of evaluation of FL performance. Foreign agencies, such as the British Council, helped with leading experts as advisors and provided training of item writers, examiners, teachers and other personnel who would be involved in the new examination process. The outcomes involved detailed requirements and test specifications, training materials and courses for examiners, item-writer guidelines, calibrated test items, books for preparation exam students etc. Unfortunately, when projects end, some good testing practices are not sustained due to a lack of financial support, appropriate infrastructure, and assessment expertise in the educational context and the wider society (Wall, 2013).

The most worrying issues may be summarized as follows:

- Absence of validation of the language tests
- Absence of piloting of test items
- Oral parts remain internal (i.e., they are delivered and rated by the students' own teacher)
- No double-marking or monitoring of rating standards in the writing and speaking tests
- Inadequate or no quality control of the test development process
- No training for novice item writers

3.5 The influence of politics

Heyneman (1987) claims that testing is a profession but may be dramatically affected by politics. He warns that the quality of tests relies on how much the test designer is willing and able to pursue professionalism in language assessment. This is even more so in the case of national FL assessments, which are usually part of a larger national educational and assessment scheme with many different agendas expressed by different stakeholders. Negotiation, compromise and concession are a major part of every test development process at this level. Furthermore, this often takes place in a complex organizational structure, with some people operating under their own particular agendas, which may be legitimate but also personal and even egoistic. They may have a completely inaccurate and/or simplistic view of what it takes to make a good test (Buck 2009:177).

At the macro-political level, national educational policy may involve innovations in assessment in order to influence the curriculum and/or teaching practice. For example, the Slovene secondary-school leaving examination was introduced as a lever for change, to promote communicative language teaching and assessment but the new language curriculum was developed and implemented only several years after the introduction of examination. Politics, however, can also operate at lower levels, and can be a very important influence on test development and its implementation. Alderson and Banerjee (2001) point out that in most

testing institutions, test development is a complex process where individual and institutional motives interact and are interwoven. Alderson (2009) further argues that politics with a small 'p' does not only include institutional politics, but also personal politics. Different stakeholders (ministers, ministry bureaucrats, university teachers, chairs of educational bodies etc.) have their own agendas and may impact the test development process and test use. Eckes et al. (2005) report that in Hungary in 2002 a few top decision-makers decided to create a unified examination model for all foreign languages despite differences of opinion between the various language teams. Classroom teachers became responsible for developing their own speaking tests, as well as marking them, with no quality control. This led the English team to resign.

3.6 Washback effect of the national foreign language assessments

Large-scale national FL assessments may have an intended and unintended impact on learning and teaching. In the research literature, this impact is referred to as the "washback effect". Most researchers define it as a complex phenomenon which influences language teachers and students to do things they would not necessarily otherwise do (Alderson and Wall 1993:117; Bailey 1996: 259). It also indicates an intended direction and function of curriculum change on aspects of teaching and learning by means of a change of an examination (Cheng 2005:28-29).

Such impact may be seen as negative; tests may be assumed to force teachers, students and other stakeholders to do things they would not otherwise do. For example, the General English Proficiency Test (GEPT) in Taiwan is targeted at high school students and adults. However, due to parents' influence, primary school students started to take the GEPT. To meet parents' expectations, language schools provided young learners with test preparation programmes. In 2006, learners at the primary school level were barred from registering for the GEPT (Wu, 2012).

On the other hand, some researchers claim that tests may also be 'levers for change' in language education: the argument being that if a bad test has negative impact, a good test should or could have positive washback (Pearson 1998). However, washback effect is a far too complex a process; any test, good or bad, may have beneficial or detrimental effects on learning and teaching. Research findings on the washback effects of the university entrance examinations on teaching and learning show that washback is inextricably linked to the context and that tests changed teachers' teaching methods in some but not all studies and that washback works on teachers at different levels (Hassan and Shih, 2013; Cheng, 2005; Qi, 2005). This implies that each examination needs a tailor-made research project to investigate its washback.

4. Future Directions

As we described earlier, national assessments have become more standardized in many countries and their design more professional. This is an indication of an increasing awareness among educational decision makers and assessment organizations that language testing needs to be taken seriously if test results are to be trusted. It is to be hoped that this trend continues and the issues with certain national assessment systems reported in, e.g., Eckes et al. (2005) will be exceptions rather than the rule in the future. Improved assessment literacy is obviously also important for language teachers for whom assessment is in fact part

of their profession (see also Inbar-Lourie 2013). Indeed, ordinary classroom teachers are one of the target groups of international language assessment associations such as EALTA in their efforts to promote a better understanding of the principles of good language testing.

A very clear trend in both national and international large scale assessments is the increasing use of computers and other types of ICT in all phases of the assessment process, from item writing to test delivery and scoring of responses. International high-stakes language examinations from the Educational Testing Service (responsible for the TOEFL) and the Pearson publishing company are the most prominent examples of very advanced utilizations of computer technology and the Internet (Chapelle, Enright and Jamieson 2008; Owen 2012). Interestingly, there are also computerized large-scale low-stakes language tests such as DIALANG, which is a multilingual diagnostic language assessment system available through the Internet that provides its users with feedback on the strengths and weaknesses in their proficiency (Alderson 2005). Large-scale programme evaluations are also becoming computerized; for example, the European Commission's recent survey of FL proficiency (European Commission 2012) was delivered on a computer in some of the participating countries. Computerization has also become the delivery mode in some national examinations (e.g., in Norway; Moe 2012) and this trend is likely to gain speed in the future.

Computerization is just not an alternative way to administer test content: it can in fact expand and change the constructs measured (see Van Moere and Downey, this volume; Sawaki 2012; Kunnan 2014). An obvious expansion is the use of multimedia in speaking and listening tests, another is the possibility of allowing test takers to use on-line dictionaries or other such tools that are used in real-life language tasks (Chapelle et al. 2008).

Increasing computerization of national assessments relates to the final trend we single out in this review, namely an increase in the amount and detail of information obtainable from such assessments. The primary purpose of most national assessments is to provide summary information for educational authorities (information about large groups of learners) or for individual learners (overall grades based on achievement). However, it is, in principle, possible to extract and report much more than just overall test scores from major tests, if this is considered useful for the stakeholders and if it is practical to do so. Recent interest in forms of assessment that support learning, such as formative (e.g. Black and Wiliam 1998), diagnostic (e.g. Alderson 2005; Alderson and Huhta 2011) and dynamic (e.g. Poehner and Lantolf 2013; Poehner and Infante, this volume) assessment has generated more interest in the value of detailed information and feedback from language assessments. Advances in the utilization of computers in testing have provided the tools to address this need in practice. The automatic calculation of sub-test and item level scores in computer based tests makes the provision of profile scores and detailed feedback a viable option for assessment organizations. Related advances in the automated analysis and evaluation of language learners' speech and writing offer truly amazing possibilities for detailed and individualized feedback to learners and their teachers (Chapelle 2008; Bernstein, Van Moere and Cheng 2010). Large-scale national assessments that provide detailed feedback to teachers and learners do not seem to exist yet. However, recent developments indicate that this may become more common in the future, such as the current plans in the Netherlands to introduce nationwide diagnostic tests in several subjects, including Dutch as L1 and English as a FL (see CITO, 2014). It is likely that many other countries will introduce similar assessment systems in the future.

5. References

Alderson, J. Charles. 2004. The shape of things to come: Will it be the normal distribution? In Michael Milanovic & Cyril J. Weir (eds.), *Studies in language testing 18: European language testing in a global context: Proceedings of the ALTE Barcelona Conference July 2001*, 1–26. Cambridge: Cambridge University Press.

Alderson, J. Charles. 2005. *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, UK: Continuum International Publishing.

Alderson, J. Charles. 2009. *The politics of language education: individuals and institutions*. Bristol: Multilingual Matters.

Alderson, J. Charles & Jayanti Banerjee. 2001. State of the Art Review: Language Testing and Assessment Part 1. *Language Teaching* 34. 213–236.

Alderson, J Charles & Ari Huhta. 2011. Diagnosing strengths and weaknesses in second and foreign language reading: What do second language acquisition and language testing have to offer? *EUROSLA Yearbook*, Vol. 11(1), 30-52.

Alderson, J. Charles & Karmen Pižorn. (eds.). 2004. *Constructing school-leaving examinations at a national level - Meeting European standards*. Ljubljana: British Council and Državni izpitni center.

Alderson, J. Charles & Dianne Wall. 1993. Does washback exist? *Applied Linguistics* 14(2). 115-129.

Anderson, Scarvia B. 1975. *Encyclopedia of educational objectives*. Jossey-Bass Publisher San Francisco. California, USA.

Bachman, Lyle F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, Lyle F. & Adrian Palmer. 1996. *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Bailey, Kathleen M. 1996. Working for washback: A review of the washback concept in language testing. *Language Testing, 13(3)*, 257–279.

Bernstein, Jared, Alistair Van Moere & Jian Cheng. 2010. Validating automated speaking tests. *Language Testing* 27(3). 355–377.

Black, Paul & Dylan Wiliam. 1998. Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan*, 80(2) 139-48.

Brown, Annie. 2013. Uses of language assessments. In Carol A. Chapelle (ed.), *The encyclopedia of applied linguistics*. Oxford, UK: Wiley-Blackwell. DOI: 10.1002/9781405198431.wbeal1237

Buck, Gary. 2009. Challenges and constraints in language test development. In J. Charles Alderson (ed.), *The politics of language education: individuals and institutions*, 166–184. Bristol: Multilingual Matters.

Chapelle, Carol A., Enright, Mary K. & Jamieson, Joan M. (eds.) 2008. *Building a validity argument for the Test of English as a Foreign Language.* New York: Routledge.

Chapelle, Carol A., Jamieson, Joan M. & Hegelheimer, Volker 2003. Validation of a Web-based ESL Test. *Language Testing*, 20(4), 409-439.

Cheng, Liying. 2005. *Changing Language Teaching Through Language Testing: A Washback Study*. Cambridge: UCLES/Cambridge University Press.

CITO. 2014. De diagnostische tussentijdse toets. Tussenstand in http://www.cito.nl/onderwijs/voortgezet%20onderwijs/diagnostische_tussentijdse_toets (accessed 1 October 2014).

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

Davies, Alan. 1990. *Principles of language testing*. Oxford: Blackwell.

Davies, Alan. 2014. Fifty years of language assessment. In Antony Kunnan (ed.), *The Companion to Language Assessment I: 1:1*, 1-21. John Wiley & Sons, Inc. DOI: 10.1002/9781118411360.wbcla127

Eckes, Thomas, M. Ellis, Vita Kalnberzina, Karmen Pižorn, Claude Springer, K Krisztina. Szollás & Constantia Tsagari. (2005). Progress and problems in reforming public language examinations in Europe: cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France and Germany. *Language Testing*, *22*(3), 355–377. DOI:10.1191/0265532205lt313oa

European Commission 2012. First European survey on language competences. Final report. http://ec.europa.eu/languages/policy/strategic-framework/documents/language-survey-final-report_en.pdf (accessed June 22, 2014)

EURYDICE. 2009. *National testing of pupils in Europe: objectives, organisation and use of results*. Brussels: Eurydice EACEA. http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/109EN.pdf (Accessed October 1, 2014).

Figueras, Neus. 2007. The CEFR, a lever for the improvement of language professionals in Europe. *The Modern Language Journal*, 91(4). 673–675. DOI:10.1111/j.1540-4781.2007.00627_8.x

Figueras, Neus. 2012. The impact of the CEFR. *ELT Journal*, 66(4). 477–485. DOI:10.1093/elt/ccs037

Fulcher, Glenn & Fred Davidson. 2009. Test architecture, test retrofit. *Language Testing*, *26*(1), 123–144. DOI:10.1177/0265532208097339.

Hassan, Nurul Huda & Shih Chih-Min. 2013. The Singapore–Cambridge General Certificate of Education Advanced-Level General Paper Examination. *Language Assessment Quarterly* 10(4). 444-451.

Heyneman, Stephen P. 1987. Uses of examinations in developing countries: selection, research, and education sector management. *International Journal of Educational Development* 7(4). 251–263.

Hughes, Arthur. 2003. *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Inbar-Lourie, Ofra. 2013. Language assessment literacy. In Carol A. Chapelle, (ed.), *The encyclopedia of applied linguistics*. Oxford, UK: Wiley-Blackwell. DOI: 10.1002/9781405198431.wbeal0605.

Kunnan, Anthony J. 2008. Large scale language assessments. In Elana Shohamy & Nancy Hornberger (eds.), *Language and Education*, Vol 7, 135-155. New York: Springer.

Kunnan, Antony J. (ed.) 2014. *The Companion to Language Assessment.* Hoboken, NJ: Wiley-Blackwell.

Martyniuk, Waldemar & Jose Noijons 2007. Executive summary of results of a survey on the use of CEFR at national level in Council of Europe member states. Intergovernmental Forum: The Common European Framework of Reference for Languages CEFR) and the development of language policies: Challenges and responsibilities. Strasbourg: Council of Europe, 6-8 February 2007, Report of the Forum. www.coe.int/t/dg4/linguistic/source/survey_cefr_2007_en.doc (Accessed October 1, 2014).

Martyniuk, Waldemar. (ed.) 2011. *Aligning tests with the CEFR: Reflections on using the Council of Europe's Draft Manual*. Cambridge: Cambridge University Press.

Moe, Eli. 2012. Valid testing of young learners – an achievable endeavour? Presentation at the ALTE Conference, Munich, November 23, 2012. http://www.alte.org/attachments/pdfs/files/valid_testing_of_young_learners_an_achievable_eneavour_eli_moe_166ml.pdf (Accessed October 1, 2014)

Owen, Nathaniel. 2012. Can PTE Academic be used as an exit test for a course of academic English? Pearson Research Notes. http://pearsonpte.com/wp-content/uploads/2014/07/Owen_Executive_Summary.pdf (Accessed October 1, 2014).

Pearson, Ian. 1988. Tests as levers for change. In Dick Chamberlain & Robert J. Baumgardner, (eds.), *ESP in the classroom: Practice and evaluation*, 98-107. Great Britain: Modern English Publications.

Pizorn, Karmen & Eli Moe. 2012. A validation study of the national assessment instruments for young English language learners in Norway and Slovenia. *CEPS journal*, 2(3). 75-

96. http://www.cepsj.si/pdfs/cepsj_2_3/cepsj_2_3_pp75_pizorn%20etal.pdf (accessed 9 February 2014)

Pizorn, Karmen & Edit Nagy. 2009. The politics of examination reform in Central Europe. In J. Charles Alderson (ed.), *The politics of language education: individuals and institutions*, 185–202. Bristol: Multilingual Matters.

Poehner, Matthew & James Lantolf. 2013. Bringing the ZPD into the equation: Capturing L2 development during Computerized Dynamic Assessment (C-DA). *Language Teaching Research* 17(3). 323–342.

Poehner, E. Matthew & Paolo Infante. This volume. Dynamic assessment in the language classroom. In Dina Tsagari & Jayanti Banerjee (eds.), *Handbook of Second Language Assessment*. Berlin: DeGruyter Mouton.

Qi, L. 2005. Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing* 22(2). 142–173.

Resnick, Lauren. B. & Daniel P. Resnick. 1992. Assessing the thinking curriculum: New tools for educational reform. In Bernard R. Gifford & Mary C. O'Connor (eds.), *Changing assessment: Alternative views of aptitude, achievement and instruction*, 37–76. London: Kluwer Academic Publishers.

Sawaki, Yasuo. 2012. Technology and language testing. In Glenn Fulcher & Fred Davidson (eds.) 2012. *The Routledge handbook of language testing* 426-437. New York: Routledge.

Tannenbaum, Richard & Elaine C. Wylie. 2008. Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology. RR-08-34. Princeton, NJ: Educational Testing Service.

Van Moere, Alistair & Ryan Downey. This volume. Technology and artificial intelligence in language assessment. In Dina Tsagari & Jayanti Banerjee (eds.), *Handbook of second language assessment*. Berlin: DeGruyter Mouton.

Wall, Dianne. 2005. *The impact of high-stakes examinations on classroom teaching.* Cambridge: UCLES/Cambridge University Press.

Wall, Dianne. 2013. Factors affecting long-term examination impact, and the fate of the examinations. Paper presented at the Tenth Annual Conference of EALTA, Istanbul, Turkey 23- 26 May. http://www.ealta.eu.org/conference/2013/programme.html (accessed 29 July 2014).

Weir, Cyril J. 2004. *Language testing and validation. An evidence-based approach.* Basingstoke: Palgrave Macmillan.

Wu, Jessica R. W. 2012. GEPT and English Language Teaching and Testing in Taiwan. *Language Assessment Quarterly*. 9(1). 11-25.

Kentän koodi muuttunut

Kentän koodi muuttunut

[1] The Eurydice Network provides information on and analyses of European education systems and policies. As from 2014 it consists of 40 national units based in 36 countries participating in the EU's Erasmus+ programme.

[2] Standard Classification of Education (ISCED) to facilitate comparisons of education statistics and indicators across countries on the basis of uniform and internationally agreed definitions. Primary education (ISCED 1) usually begins at ages five, six or seven and lasts for four to six years. Lower secondary education (ISCED 2) generally continues the basic programmes of the primary level, although teaching is typically more subject-focused, often employing more specialised teachers.