Ari Huhta, Gudrun Erickson, Neus Figueras (eds.)

Developments in Language Education: A Memorial Volume in Honour of Sauli Takala

EALTA – European Association for Language Testing and Assessment
University of Jyväskylä, Centre for Applied Language Studies

Developments in Language Education:
A Memorial Volume in Honour of
Sauli Takala

Developments in Language Education: A Memorial Volume in Honour of Sauli Takala

Edited by
Ari Huhta, Gudrun Erickson, Neus Figueras

University of Jyväskylä/Centre for Applied Language Studies
EALTA-European Association for Language Teaching and Assessment
Jyväskylä & EALTA, 2019

Cover photograph (Köklot Island, Kvarken Strait) and the photograph of Sauli Takala by Paula Bertell
Cover and layout Sinikka Lampinen

ISBN 978-951-39-7748-1 (PDF)
ISBN 978-951-39-7747-4 (pbk)

University Printing House, Jyväskylä, Finland 2019

Table of contents

Introduction

This Volume implements the decision taken at EALTA's Annual General Meeting held in Sèvres, France, on 2nd June 2017, to honour the memory of Professor Sauli Takala, a founding member of the Association, its President from 2007 to 2010, who tragically left us on 15th February 2017.

Sauli was emeritus professor at the University of Jyväskylä, Doctor of Philosophy honoris causa and Doctor of Education honoris causa. He was committed to the values of the Council of Europe and was heavily involved in the Council's developments in language education over decades. Although retired from his professorship in applied linguistics in 2002, this may have gone unnoticed given his tireless dedication to the field of applied linguistics and specifically language testing and assessment. During his extremely busy retirement years, he took part in many projects and academic evaluations such as being the public opponent for numerous licentiate and doctoral theses. He was the recipient of a number of awards and honorary doctorates.

Sauli was a great friend, mentor and colleague, and those of us who had the pleasure of working with him, enjoyed his scholarship, his gentle attitude and his genuine love for the fields of language teaching and assessment. He made a tremendous contribution to the assessment community in his unique considerate way, drawing on his wealth of knowledge and experience. He was always active for language assessment professionals world-wide, always willing to share his wisdom and his vast collection of books, materials and articles (see further https://kiesplang.fi). Sauli had an open mind, a warm heart and a winning personality.

This memorial volume comprises an impressive 23 texts, with considerable variation thematically and in structure, as well as regarding length and genre. All, implicitly or explicitly, reflect Sauli Takala's wide and diverse areas of expertise and interest and also his many co-operations and networks. Abstracts in English are provided for texts that are written in Finnish or Swedish. The contributions are placed in alphabetical order according to the surname of the author(s).

A digital copy of this volume is available in the Resources section of the EALTA webpage (http://www.ealta.eu.org/resources.htm).

We as editors, and also on behalf of EALTA, are extremely grateful to the authors of the different contributions and chapters, to our co-editors Kathryn Brennan and Elizabeth Guerin, and to all those professionals at the University of Jyväskylä who made this Volume possible with their work and expertise.

15 April, 2019

Ari Huhta, Jyväskylä
Gudrun Erickson, Gothenburg
Neus Figueras, Barcelona

Linking vocabulary to the CEFR and the Global Scale of English: A psychometric model

Veronica Benigno

Pearson, Germany

John de Jong

Language Testing Services, The Netherlands

1. Introduction

A review of the studies on vocabulary acquisition, teaching and assessment (Bogaards and Laufer, 2004; Granger, 2017; Meara, 2009; Milton, 2009; Read, 2000; Schmitt and McCarthy, 1997) shows that although this field of investigation has evolved quite rapidly over the last few decades, there is still little agreement on how many and which words are needed to communicate efficiently at increasing proficiency levels. Attempts to relate vocabulary knowledge to proficiency levels have focused on quantitative aspects and frequency profiling methods have investigated learners' knowledge of single words, for example, by means of vocabulary size tests, e.g. the *Vocabulary Levels Test* (Nation, 1990; Schmitt, Schmitt, and Clapham, 2001). In reality, vocabulary knowledge is known to be a multidimensional construct (Daller et al., 2007; Meara, 2009; Read, 2004) which should, therefore, not simply be regarded as a quantitative process (in terms of expansion of one's vocabulary size) but also be considered from a qualitative point of view (in terms of vocabulary depth, e.g. knowledge of different word meanings, collocations, pragmatic rules).

At present there is no clear guidance on vocabulary requirements at different proficiency levels. On the contrary, several recent studies show that the selection of vocabulary included in textbooks lacks methodological foundation. Norberg and Nordlund (2018) analyse a corpus consisting of seven textbooks commonly used in Swedish primary school years 3 and 4 (students aged 9–10 years). They compare it with the New General Service List (NGSL, Browne, 2013) and the VP-Kids corpus (Roessingh and Elgie, 2009) and find that Swedish books contain a high proportion of lexical words seldom used by native-speaking children, while covering only a limited set of the NGSL. A similar study was carried out in the Netherlands (De Jong, 1989) showing that the 9 most used coursebooks for the then recently introduced subject English in primary schools together contained almost 10 thousand different words but that only 100 words occurred in all nine coursebooks. Obviously, this creates a problem

for teachers in secondary schools as young children from different primary schools, having been taught from different coursebooks, each know very different sets of words.

The CEFR publication (Council of Europe, 2001) provides only a vague definition of vocabulary knowledge, indicating, for example, that "a very broad lexical repertoire" is needed at the C2 level. It should be noted, however, that the CEFR explicitly (Council of Europe, 2001, 23) builds onto the earlier work in language learning of the Council of Europe and that the language exponents for the levels A1 to B2, including the lexical elements, have been thoroughly described in the corresponding publications Breakthrough (Trim, 2009), Waystage (Van Ek, Alexander, and Fitzpatrick, 1977; Van Ek and Trim, 1990), Threshold Level (Van Ek, 1975, Van Ek and Trim, 1990), and Vantage (Van Ek and Trim, 1996). Nevertheless, the Council of Europe made the recommendation to the state members to create, for each regional and national language, inventories of linguistic forms known as *Reference Level Descriptions* (Council of Europe, 2005). The RLDs are "inventories of the linguistic realisations of general notions, acts of discourse and specific notions/ lexical elements and morpho-syntactic elements" which are characteristic of each level (Council of Europe, 2005, p.5).

One approach to inventorying vocabulary by CEFR level is based on learner corpora: see, for example, Cambridge *Vocabulary English Profile* (www.english-profile.org/ wordlists). The rationale is that in order to determine which words (word meanings) are needed to perform at a certain CEFR level a learner corpus can be compiled from responses to questions in an exam that is claimed to be at that CEFR level. We believe there are two major issues with this approach. First, exams can only deal with limited samples of language and thereby, as it where, steer the language that the examinee is bound to use to deal with the exam questions. In a research study analysing 28,320 test taker responses O'Loughlin (2013) found that the variation in percentage of academic words (from the Academic Word List (AWL), Coxhead, 2000) in item prompts explained 80% of the variance in academic word usage in the responses. Secondly, assigning the vocabulary usage found in responses to an exam to the level of the exam and consequently to a CEFR level, suggests that the exam is indeed a valid operationalisation of that CEFR level.

Another approach has recently been operationalised by Brysbaert et al. (in press) by introducing the concept of word prevalence, which indicates how many people know a word. The researchers obtained their measure on the basis of an online crowdsourcing study involving over 220,000 people and identifying what words are most commonly used (and therefore known) by a large sample of native speakers. A drawback with this approach is that it deals only with word knowledge at the level of lemmas and ignores the meaning. The word '*lot'*, for example, figures as a lemma known by more than 99% of all people in the sample; but '*lot'* has many meanings and the usage of the lemma does not necessarily mean that all people using the lemma '*lot*' in a phrase like *'I like it a lot'* will also be aware of its meaning as 'an area of land' or 'quality/conditions of life'.

The present study, set up in response to the Council of Europe recommendation (Council of Europe, 2005) mentioned above, produced a vocabulary database of about 37,000 word meanings represented by 20,000 lemmas, each linked to a CEFR level and a GSE value. The database is freely made available online by Pearson at https://www.pearson.com/english/about/gse/teacher-toolkit.html and aims to help users select graded vocabulary, according to the targeted proficiency level. In section 2 we present the theoretical background to the study. In section 3 we briefly describe how the vocabulary database was compiled. Section 4 describes the statistical analysis performed to scale vocabulary against proficiency. Section 5 briefly discusses the contribution of the study to the fields of language teaching and assessment.

2. Word meaning, word frequency, and word usefulness

In recent years, many studies have shown that language is formulaic with no rigid separation between vocabulary and grammar (Ellis, 2002; Wray, 2002). Words occur most frequently in a limited number of contexts (i.e. in co-occurrence with a limited number of other words), producing collocations, chunks, ready-made phrases, and fixed units. Other research has shown that partial acquisition of a word meaning is a very common stage in language development since learners encounter and use words in a number of predictable lexical environments and gradually extend their knowledge as their proficiency increases (Wolter, 2009). Research on vocabulary size has mainly used two units of count to describe how large someone's vocabulary is: the lemma and the word family. These have been operationalised to assess vocabulary size, or to measure lexical coverage, or to produce word lists for pedagogical purposes. Counting by lemmas or word family is methodologically easy because it makes use of readymade lists where the main word and its inflected forms or family members respectively are listed but comes with a few theoretical issues. If we took the word-family as unit of count, we could think it implies that if a learner knows the main member (e.g. the base word *row*) of a word family, then he/she would know all other members inflected and derived forms) within the same family (e.g. the noun and the verb *row* and related words such as *rows, rowed, rowing*). However, research (Schmitt and Zimmerman, 2002; De Jong 2002) has provided counterevidence of this assumption, suggesting alternative units of count such as the lemma. Furthermore, the concept of word family is not rigorously defined. For example, the lemmas *abbes, abbey and abbot* are not in the same word family, whereas *act, actor,* and *actress* together with even *action* and *inaction* do belong to one and the same family. Counting by lemmas, on the other hand, also produces a limited view of vocabulary knowledge as it does not help distinguish between different parts of speech and word meanings, e.g. between *row* in the meaning of "people or things in a straight line" and *row* in the meaning of "dispute". Since vocabulary learning takes place in context and different meanings of polysemous words are most likely learned at different stages of proficiency (Anderson and Freebody, 1979), our study claims that the best unit of count to express someone's vocabulary

knowledge is the number of word meanings he/she knows. Therefore, in the present study, each learning unit is a word meaning, not a lemma (a base word form and its inflected forms within the same part of speech) or a word family (a word and its related inflections and derived words).

Another drawback of current research on vocabulary measurement is the use of frequency of occurrence (in a reference corpus) as main criterion to establish a rank between lexical units. Although many studies outline the importance of frequency of exposure as an objective criterion in deciding what to teach first (Ellis, 2002; Gyllstad, 2007; Nation and Beglar, 2007), frequency alone is not sufficient to identify pedagogically-relevant vocabulary. According to Widdowson (2003, p.83), "[…] prototypical prominence in the mind does not accord with frequency of actual occurrence". Milton and Alexiou (2009, p. 198) ascertain the relevance of frequency yet acknowledge the influence of other factors. A purely frequency-based pedagogical list is necessarily biased by the nature of the corpus and would ignore low-frequency words which refer to basic concepts that are useful for communicative purposes but rarely spoken or written about by users of the language. As Stubbs (2002) points out, the definition of what is basic depends not only on frequency, but also on functional criteria such as communicative relevance or usefulness. Therefore, the present study combines the two criteria of word frequency (retrieved by corpus analysis) and word usefulness (derived from teacher ratings) to determine which word meanings should be learned first.

3. Compilation of the database

A database of about 37,000 word meanings (20,000 lemmas) was compiled in a number of steps requiring both automatic and manual analysis. Each (word meaning) entry in the database was assigned a part of speech, a dictionary definition, a topic and subtopic tag, collocations and phrases, a frequency value and a usefulness value. In a final step reported in section 4, each word meaning was also assigned a CEFR level and a GSE value. Below are the main steps followed to compile the database.

a. A lemma-based frequency list was extracted from L1 reference corpora. The chosen reference corpora included spoken and written texts of general English. They were the *Longman Corpus Network (*http://www. pearsonlongman.com/ dictionaries/corpus/), of about 330 million tokens; *UKWac* (Baroni et al., 2012), a web-crawled corpus of about 2 billion+ tokens; and the spoken section of the *Corpus of Contemporary American English* (http://corpus.byu.edu/coca/), of about 90 million tokens. It was chosen to extract the top 10k items in line with research on vocabulary size claiming that about 10,000 learning units is the required target to successfully communicate in another language at the B2 CEFR level (e.g. Hazenberg and Hulstijn, 1996; Laufer and Nation, 1999)

b. The extracted list was filtered and refined. First the list was cleaned from any "noise" deriving from the automatic extraction, e.g. spelling errors producing incorrect lemmatisation or part-of-speech tagging. Then the Pearson *Longman Active Study Dictionary of English* (Pearson, 2000), which counts about 25,000 entries, was consulted to expand or refine the list with new entries, obtaining a final list of 20,000 lemmas corresponding to about 37,000 word meanings.
c. The next step was to link each lemma to its dictionary definition(s). In this way, polysemous lemmas were disambiguated and multiple entries were created in the database, each corresponding to a word meaning (instead of a lemma). It should be noted that each word meaning of polysemous words carried the same frequency value of the corresponding lemma because the corpus analyses do not provide different word meaning frequencies. At this stage, each word meaning was assigned a topic and subtopic tag following the Council of Europe categorisation in Specific Notions, General Notions, and Functions included in the *Vantage* Specifications (van Ek and Trim, 2001). In this way, each word meaning was assigned a topic and subtopic tag, e.g. "food and drinks", "sport", "body and health", "science and technology". Moreover, about 80,000 collocations and 7,000 functional units/pragmatic phrases were also added under each word meaning to provide additional context.
d. Each entry (word meaning) was presented to a pool of 20 EFL teachers who were asked to rate the importance of each word meaning separately in order to produce usefulness values. This rating exercise was carried out to be able to rank the word meanings which, by frequency, were assigned exactly the same rank. Teachers received online training and followed specific guidelines. The underlying principle of the rating exercise was the one of efficiency: What vocabulary gives learners the highest chance of communicating with other speakers? What is the relative importance of vocabulary items to be able to participate in a general conversation? Each word was rated by a random 10 out of the 20 raters using a pre-defined scale of usefulness going from value 1 (=essential) to value 5 (=extra). In addition, raters could choose not to rate a word by assigning the (arbitrary) value 99. This was the case if they had never heard of the word before or they could not decide between widely different ratings. Following the data cleaning (see section 4), one rater was discarded due to low intra and inter reliability of his ratings compared to the group.
e. In a final step, which is the main focus of the present paper (reported in section 4), frequency and usefulness values were combined to produce a weighted measure to link each word meaning to proficiency on the CEFR and the Global Scale of English.

More details about the methodology used to compile the vocabulary database as well as the available features are in Benigno and De Jong (2017) accessible at https://prodengcom.s3.amazonaws.com/GSE-Vocab.pdf.

4. Statistical analysis

In this section we describe the statistical analysis carried out to scale vocabulary against proficiency. The analysis consisted of three main steps outlined in the subsections below:

- Data cleaning (subsection 4.1): to evaluate the soundness of the rating data and remove unreliable ratings
- Combined analysis of frequency and rating data (subsection 4.2): to produce a formula to rank vocabulary using the collected frequency and rating values
- Data modelling (subsection 4.3): to fit the data onto a model of vocabulary learning which is in line with the current research evidence on vocabulary size.

4.1 Data cleaning

This step consisted of the analysis and cleaning of the collected teacher ratings. As outlined in the previous section, each entry (word meaning) was presented to a pool of 20 EFL teachers. By means of an overlapping design, each entry was randomly assigned for rating by 10 out of these 20 teachers. All teachers had English as their L1. Before starting the rating exercise, teachers attended an online standardisation session and received written guidelines. Their task was to rate the importance of each word meaning separately in order to produce usefulness values. They were instructed to use the following pre-defined scale of usefulness and asked to assign a level of relevance to each word meaning by choosing only one value between: 1, 2, 3, 4, 5, and 99:

- 1 *"Essential" (words learners would want to acquire first)*
- 2 *"Important" (words that become necessary at a next stage)*
- *3 "Useful" (words enabling more detailed and specific language*
- *4 "Nice to have" (words to express concepts more accurately)*
- *5 "Extra" (words some language users will use occasionally)*
- *99 "Escape" (words which are impossible to rate - you have never heard of the word before or you cannot decide between widely different ratings).*

Teachers were asked to use their common sense, knowledge of the language, and expertise as teachers to inform their rating decisions, evaluating how useful each entry was for general communication. They were sent rating batches of about 500/600 entries in a spreadsheet: each line showed a word meaning and its part of speech, definition, and example sentence (when available in the Pearson Longman dictionary database).

Most importantly, teachers were rating meaning within the same topics. For example, the word "row", which is highly polysemous, was presented in as many different batches as the number of topics in which it appears (as shown in Figure 1 below) – in order to facilitate the raters' task to rank different word meanings of a same lemma.

Figure 1: Occurrence of the lemma "row" in different rating batches based on different meanings and contexts.

Rating batch	Topic	Headword	POS	Definition	Example
1	Holidays, travel, and transportation	row	verb	to make a boat move across water, using oars	They rowed across the lake
2	Interacting with others	row	noun	an argument	Anna and her boyfriend are having another row
2	Interacting with others	row	noun	a situation in which people disagree strongly about important public affairs	a row over government cuts
2	Interacting with others	row	verb	to argue in an angry way	They rowed about money all the time
3	Media, arts, literature, and entertainment	row	noun	a line of seats in a theatre, cinema etc	Gabrielle found a seat in the front row
4	Physical attributes	row	noun	an annoying loud noise	*
5	Politics and society	row	noun	a situation in which people disagree strongly about important public affairs	a row over government cuts
6	Quantity or number	row	noun	a line of things or people next to each other	a row of houses
7	Sports, hobbies, and interests	row	verb	to make a boat move across water, using oars	They rowed across the lake

*Word meaning definitions, part of speech and example sentences for each lemma were extracted from the Pearson Longman dictionary database via a fully automatic process. In a dictionary, sentence examples are provided only if they are useful to illustrate the meaning of a word more clearly than is possible by the sole definition. Therefore, many entries shown to our raters were lacking an example sentence, as is the case for "row" defined as "an annoying loud noise". However, the feedback we collected from teachers showed they had no issue recognising it as a separate meaning.

The original dataset consisted of a total of 372,265 teacher ratings, i.e., on average about one rating assigned by 10 out of the 20 teachers to each of the 37,214 word

meanings. From this dataset ratings were removed according to the criteria below, reducing the data size to 349,861 ratings.

- All “99” ratings, i.e., the code given by teachers to indicate they were unable to assign a usefulness rating from 1 to 5, were removed and replaced by empty cells, effectively interpreted as missing data.
- Misfitting ratings were then removed too. A misfitting rating is any rating in the range 1 to 5 which was more than 1.5 points distant from the group average rating for a particular entry (word meaning). After removal of 22,404 ratings in total (Table 1 and Table 2), the data set reduced from 372,265 (100%) to 349,555 (94%). Among the 22,404 (6%) removed ratings, 5,326 (1.4%) were “99” ratings and 17,384 (4.7%) were misfitting ratings. Below we will explain this procedure more in detail.

First, we removed all “99” ratings provided by the teachers, i.e., 5,326 ratings. The “99” ratings were not included in the statistical analysis because they indicated that raters could not provide a usefulness judgement. The raters’ use of “99” ratings ranged from 0% to 20.4%, with an average of 2% and a standard deviation of 5% (see table 1). After removal of all “99” ratings, the data size decreased to 366,939 ratings.

Table 1. Total ratings collected and removal of “99” ratings.

	All raters	Average/rater	StDev
Total Ratings	372,265	18,613	8,584
Count "99"	5,326	266	408
Remaining ratings	366,939	18,347	8,533

Next, we removed misfitting ratings, in total 17,384 ratings, i.e. 4.7%. Misfit of individual ratings was defined as a distance from the average rating greater than 1.5. In fact, we aimed to clean the data set from any misfitting data point while keeping as many ratings as possible and decided, therefore, in a first round not to remove one or more complete raters with poor statistics, but to remove individual (misfitting) ratings. The raters’ misfitting ratings based on this definition ranged from less than 1% to 51%, with an average of 17% and a standard deviation of 16%. After removal of all “99” ratings and all misfitting ratings, the data size decreased to 349,555 ratings in total (see table 2).

To check for any bias introduced by the removal of misfitting ratings, we described the data set before and after the cleaning using three different measurements: Inter-rater reliability; Intra-rater reliability; and Range.

Table 2. Removal of misfitting ratings.

	All raters	Average/rater	StDev
Total Ratings (-"99")	366,939	18,347	8,533
Count Misfits	17,384	869	813
Remaining ratings	349,555	17,478	8,425

The inter-rater reliability measured the extent to which individual rater's behaviour aligned to the average behaviour of the overall group of raters. It was computed by measuring:

- Overall average and SD of each rater's mean compared with all other raters
- Overall average and SD of each rater's SD compared with all other raters
- Overall average and SD of each rater's correlation (between individual ratings and group ratings for each entry) compared with all other raters.

Removal of misfitting ratings had little effect on the mean ratings (from 3.52 to 3.58) and on the average standard deviation of the individual raters (from 1.08 to 1.06 but did somewhat reduce the variance among raters (from 0.54 to 0.42) and improved the average of each rater's correlation with all other raters (from 0.77 to 0.84) – as shown in table 3.

Table 3. Impact of removal of individual ratings.

	BEFORE CLEANING	**AFTER CLEANING**
Average of all raters' means	3.52	3.58
SD of all raters' means	0.54	0.42
Average of all raters' SDs	1.08	1.06
SD of all raters' SDs	0.16	0.16
Average of all raters' correlations	**0.77**	**0.84**
SD of all raters' correlations	0.09	0.03

The above analysis highlighted one rater with a significantly low mean (z= -2.93) as highlighted by the standardised rating computed with Fisher's z-score $p<0.01$. This same rater had a highly significant low correlation (z = -3.33) – as shown in table 4 below. It was also this same rater that had 51% of misfitting ratings mentioned above.

Table 4. Individual raters' standardised mean, standard deviation and correlation.

Rater Code	CB	CE	CR	CT	FO	HB	HC	ID	JG	JO	OMW	SB	SW	TA	HS	KE	KM	MJH	JC	NG
Z rater mean	0.32	0.45	-0.76	-0.27	0.16	0.38	1.20	1.29	-0.34	0.01	-0.44	0.85	-0.30	-1.20	1.28	0.51	-1.06	0.80	0.05	-2.93
z rater stDev	-1.20	-1.36	-0.97	0.40	-0.60	0.11	0.19	-1.93	1.20	1.62	0.06	-0.70	-0.27	0.61	-0.75	-0.24	0.16	1.32	0.79	1.57
z Rater Correlation	-1.37	-0.08	0.02	0.44	0.56	0.69	0.55	-0.14	-0.80	0.74	-0.18	1.07	0.53	-0.29	0.37	0.65	0.23	-0.65	0.98	-3.33

The intra-rater reliability measured whether the distribution of the 1 to 5 ratings used by each rater deviated from the average distribution used by all raters. It was computed by measuring the usage of scale points and average and standard deviation obtained by each rater.

Removal of misfitting ratings slightly decreased the average number of "1" and "2" ratings, whereas it slightly increased the average number of "3", "4", and "5" ratings. Table 5 below shows these shifts in percentage.

Table 5. Impact of data cleaning on ratings usage.

	BEFORE CLEANING		**AFTER CLEANING**	
Rating	Average	Stdev	Average	Stdev
1	8%	12%	6%	7%
2	11%	5%	10%	5%
3	24%	9%	26%	9%
4	32%	10%	34%	9%
5	23%	14%	24%	14%

Next, we analysed the percentual usage of the 5 rating points by each of the raters and in how much their usage deviated from the average usage by all raters (using Fischer's z). This analysis highlighted that rater NG showed significant overuse of rating "1" (35% vs. an average of 6% by all raters) resulting in a Fisher's z-score of 3.87. As this same rater was also flagged as deviant by their percentage of misfitting ratings (51%), their average rating and by the inter-rater reliability measure (see Table 4), it was decided to remove this rater. Furthermore, rater CR had significant overuse of rating "3" (45% vs. the average usage of 26%, z = 2.14). As this rater had below average removal because of misfitting ratings and had not shown significant deviance in their average ratings and their correlation with the other raters, we decided their ratings could still be used. Table 6 provides the details on the raters' usage of the ratings.

Table 6. Individual raters' usage of rating points after data cleaning.

Rater	CB	CE	CR	CT	FO	HB	HC	ID	JG	JO	OMW	SB	SW	TA	HS	KE	KM	MJH	JC	NG	Average	Stdev
Rating	Percentual usage of rating categories, by rater																					
1	2%	1%	6%	7%	4%	4%	4%	0%	7%	11%	6%	2%	5%	11%	2%	3%	9%	6%	7%	35%	6%	7%
2	8%	5%	12%	13%	8%	9%	6%	2%	16%	12%	13%	6%	12%	17%	5%	9%	19%	7%	11%	16%	10%	5%
3	29%	30%	45%	29%	27%	23%	13%	14%	31%	22%	32%	18%	34%	37%	14%	23%	36%	14%	24%	18%	26%	9%
4	48%	45%	35%	34%	45%	36%	26%	44%	22%	25%	39%	42%	38%	25%	32%	38%	29%	24%	32%	18%	34%	9%
5	14%	19%	3%	17%	16%	27%	51%	40%	24%	30%	11%	33%	12%	10%	47%	28%	7%	49%	25%	14%	24%	14%
Rating	Z-value of percentual usage of rating categories (indicates individual raters' deviation from the average)																					
1	-0.67	-0.74	-0.09	0.02	-0.36	-0.34	-0.32	-0.85	0.07	0.55	-0.12	-0.59	-0.16	0.62	-0.65	-0.52	0.35	-0.12	0.06	3.87		
2	-0.58	-1.14	0.38	0.63	-0.49	-0.17	-0.88	-1.84	1.20	0.42	0.51	-0.99	0.37	1.55	-1.16	-0.33	1.83	-0.81	0.27	1.21		
3	0.41	0.50	2.14	0.40	0.16	-0.28	-1.46	-1.27	0.58	-0.37	0.69	-0.92	0.87	1.32	-1.36	-0.30	1.19	-1.26	-0.17	-0.87		
4	1.60	1.25	0.10	0.05	1.35	0.31	-0.87	1.18	-1.34	-1.05	0.56	0.96	0.45	-1.06	-0.16	0.44	-0.61	-1.14	-0.18	-1.85		
5	-0.69	-0.32	-1.47	-0.50	-0.57	0.22	1.88	1.10	0.03	0.45	-0.87	0.61	-0.85	-0.99	1.65	0.30	-1.14	1.79	0.10	-0.72		

Finally, the range of the ratings per entry (word meaning) provides an indication of the degree of agreement among the raters. It was computed by counting the difference

between the maximum and the minimum rating in the set of all ratings for each entry. With values 1 to 5 to choose from, the minimum possible range is 0 (all raters agree), and the maximum possible range is 4 whenever one rater opts for a rating of 1 and one other rater opts for 5. A range of 0 (zero) indicates perfect agreement among all raters involved. A range of 1 or 2 still indicates fairly good agreement between the raters, whereas a range of 3 or 4 is an indication of uncertainty among the raters. Before removing unreliable ratings, 65% of the total number of ratings had a range below 3. After the data cleaning, only 3% of the ratings had a range of 3, no range above 3 was observed, therefore, 97% was at or below 2 (see Table 7).

Table 7. Impact of data cleaning on observed rating ranges per word meaning.

	BEFORE CLEANING		**AFTER CLEANING**	
Range	Freq.	Cum%	Freq.	Cum%
0	234	1%	409	1%
1	6,550	18%	9456	26%
2	17,393	65%	26,306	97%
3	10,174	92%	1,056	100%
4	2,876	100%	0	100%

4.2. Producing a weighted formula

After cleaning the data from unreliable ratings as described in the previous section and computing the average rating from the remaining data, the next step was to create a formula to combine the frequency and the rating information to rank word meanings from the most useful to the least useful item. The following actions were taken:

- Frequency values were linearly rescaled to decimal values between 1 and 5 to enable a more direct and transparent comparison with the usefulness ratings
- The reliability (certainty) of the ratings was calculated
- A formula was developed parameterising both frequency data and rating information to rank vocabulary (from the most useful to the least useful word meaning)
- Modelling and regression analysis were performed to transform the ranking into GSE values.

4.2.1 Rescaling frequency

The frequency data ranged from 51,891 to 0 (per 1 million words). In order to project the frequency data on a similar scale as the ratings, frequency values were first transformed into absolute rankings ranging from 1 to 54,590, the most frequent entry being assigned 1 and the least frequent 54,590. These rankings were normalised using Fischer's z transformation. The z values were then transformed by linear regression to a scale with a minimum of 1.0 (representing the highest frequency observed in the data)

to a maximum of 5.0 (representing the lowest frequency observed). It should be noted that different word meanings of the same lemma will have the same frequency value in the corpus, the same z-value and the same value on the 1-5 scale.

4.2.2 Measuring reliability of ratings

Next, the reliability of the usefulness ratings was calculated. In rating research where raters rate on a categorical scale which in fact reflects an underlying continuous scale, the degree of agreement amongst raters can be expressed as the largest proportion of ratings in maximally two adjacent categories and expressed as a value from 0 to 1. In our study, we measured reliability by adding the proportions in the pair of adjacent categories that together produced the highest value of all possible pairs of adjacent categories. We choose two adjacent categories based on the assumption that ratings which differ more than one point show disagreement among the raters, whereas a one-point difference between two ratings simply means that the rated object is seen by the raters to be in proximity of the border between two categories and does not, therefore, imply relevant disagreement. If, for example, given 100 ratings assigned to a word meaning, the proportion of ratings is 0.50 for category "2", 0.13 for category "3" rating, and 0.37 for "4" rating, then the obtained certainty value is of 0.63; whereas if, given 100 ratings, the proportion of ratings is 0.50 for "2" rating, 0.37 for "3" rating, and 0.13 for "4" rating, then the certainty value is of 0.87 (see table 8 below).

Table 8. Computation of certainty (reliability) values.

Entry	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5	Certainty*
entry 1	0.00	0.50	0.13	0.37	0.00	0.63
entry 2	0.00	0.50	0.37	0.13	0.00	0.87

* Certainty is computed by adding the proportions in the two adjacent categories that together produce the highest value

All items with reliability values below 0.7 were flagged. A total of 1,813 (5%) items had reliability values below 0.7. However, the data cleaning previously performed resulted in removing deviant ratings thereby reducing the number of available ratings per word meaning. As a result, not all entries had the required number of 10 ratings. Removed ratings were eliminated from the calculations because of their deviance from the average ratings. Therefore, the certainty values of items with less than ten ratings were in fact inflated as less than ten raters would have agreed closely to the final average. Therefore, to adjust the reliability values for the number of raters involved, we developed a formula to adjust the observed reliability values for the actual number of ratings available. Reliability values were adjusted using the following formula:

$r^* = (r/16)(n+6)$

Where "r*" is the weighted certainty value; "r" is the obtained certainty value and "n" is the number of available ratings. As can be readily seen, the application of this

formula when 10 ratings are available will have no effect, because after dividing the observed certainty value by 16 it is subsequently multiplied by the same number (10+6). Yet, when only 5 ratings are available, even a perfect agreement among these remaining 5 ratings within two adjacent categories will be reduced to a value below minimum acceptability: because (5+6)/16 = 0.6875. The rationale is that having just 5 remaining ratings implies that half of the ratings deviated more than 1.5 from the mean. Clearly no matter how much the remaining five ratings agree, there is too much uncertainty. The effect of the formula is that the minimum reliability value of 0.70 is acceptable only if 10 ratings are available. If fewer ratings are available, then the threshold for the reliability value is set higher by this formula. Application of the formula in effect requires increasing the minimally acceptable observed reliability values as the number of ratings is less than 10 (see table 9).

Table 9. Minimally acceptable certainty with different numbers of raters.

N raters	Minimal Certainty
10	0.70
9	0.75
8	0.80
7	0.86
6	0.93
≤5	Never

4.3 Combining frequency and ratings

After rescaling the frequency and adjusting the ratings for reliability, a formula was used to combine frequency and ratings into one value that ranks vocabulary from the most useful to the least useful word meaning. In combining the frequency and the rating data, the weights of the rating values were adjusted based on the rating reliability. The following formula was used:

$$Combine = (1x\ FreqRank + (1 - R_{Rating})\ x\ FreqRank + R_{Rating}\ x\ RatingAvg)/2 \quad \text{(Formula 1)},$$

where

Combine is the optimal combination of ratings and frequency data, *FreqRank* is the scaled frequency rank, e.g., 2 (see subsection 4.2.1), R_{Rating} is the reliability (= weighted certainty value) of the rating data, e.g., 1 (see subsection 4.2.2), *RatingAvg* is the rating average, e.g., 2.5. If for example a word was rated by 10 raters and half of them gave a rating of 3 whereas the other half a rating of 2, then the rating average is 2.5.

If the reliability of the rating average equals 1, then the frequency rank and the rating average have equal weight in computing the combined value: their sum is divided by 2. If the reliability of the rating is <1, the rating average has the weight of its

reliability value and the frequency rank is weighted for 1+(1-reliability value) and the resulting sum is divided by 2, as outlined in table 10.

Table 10. Examples of values resulting from combining frequency and rating data.

Scaled Frequency	Rating Average	Rating Reliability	Combined value
2	2	1.00	2.00
2	2	0.90	2.00
2	2	0.80	2.00
2	2	0.70	2.00
3	2	1.00	2.50
3	2	0.90	2.55
3	2	0.80	2.60
3	2	0.70	2.65
1	2	1.00	1.50
1	2	0.90	1.45
1	2	0.80	1.40
1	2	0.70	1.35

After producing a combined value, it was decided to run an additional check to flag any suspicious items using the following four measures:

- The number of available ratings. Any entry with less than 6 ratings was flagged.
- The reliability value: any entry with a certainty value lower than 0.70 was flagged.
- The distance between the combined rank value and the average rating for each entry: any entry where the difference between the combined rank value and the average rating was larger than two standard deviations (i.e. outside the z-score range -1.96 to +1.96) was also flagged.
- The distance between the combined rank value and the frequency rank for each entry: any entry where the difference between the combined rank value and the frequency rank was larger than two standard deviations was also flagged.

A total of 4,496 entries (i.e., 1.2% of all entries) were flagged by one of the above measures and, therefore, manually checked by three raters who were asked to decide whether the flag could be ignored or instead considered to signal a genuine issue. In the latter case the entry was sent for rating again. Among the three raters one acted as adjudicator in case of disagreement between the other two raters.

4.4 Transformation to CEFR levels and GSE scores

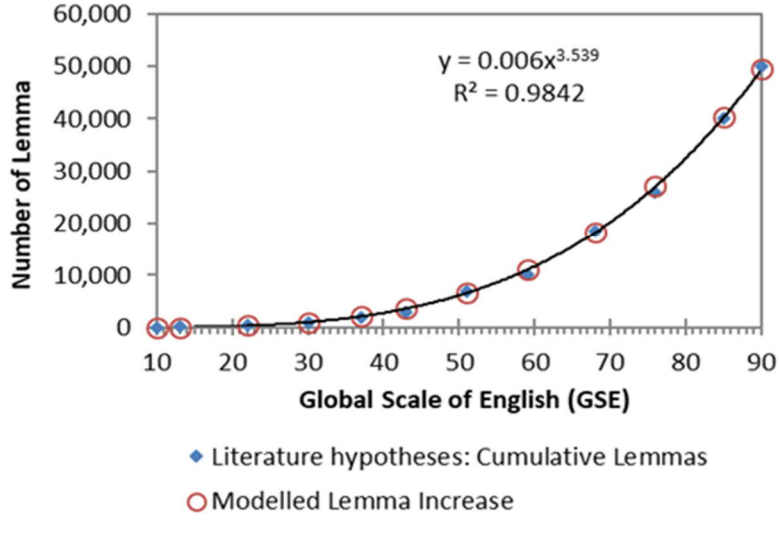
The procedure for deriving the formula can be summarised as follows. Based on the literature (e.g., Van Ek, 1975; Van Ek, Alexander and Fitzpatrick, 1977; Laufer and Nation, 1995; Hazenberg, and Hulstijn, 1996; Milton and Alexiou, 2009) on vocabulary development a total number of lemmas acquired at each of the CEFR levels is hypothesised. The lower cut-offs of these CEFR levels correspond to values on the Global Scale of English.

Table 11. Modelling vocabulary growth.

CEFR	GSE	CumLem	ModelLem
Start	10	10	21
Tourist	13	100	53
A1	22	500	338
A2	30	1,000	1,013
A2+	37	2,000	2,128
B1	43	3,000	3,622
B1+	51	7,000	6,626
B2	59	10,000	11,097
B2+	68	18,500	18,340
C1	76	26,000	27,186
C2	85	40,000	40,398
Finish	90	50,000	49,456

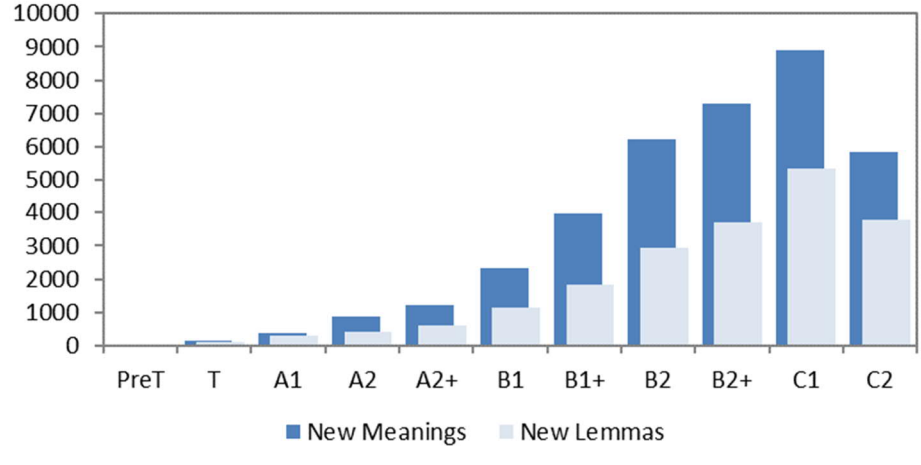
CumLem: Cumulative Lemmas FROM Literature hypotheses
ModelLem: Mathematical model to reflect hypotheses

Figure 2: From GSE to Modelled Lemma Increase



We derived a mathematical model from plotting the hypothesised lemma growth against the GSE (Figure 2) and the number of new lemmas by CEFR level can be deducted from this model. The model explained more than 98% of the variance over the CEFR levels. The data from the literature are theoretical and assume a theoretically infinite corpus. The corpus we used obviously is not infinite: it contains a little over 20,200 lemmas which represents 37,214 word meanings. The ratio of the number of lemmas by CEFR level was then used to estimate the number of lemmas per CEFR level within the limited corpus we used. The corpus contains many polysemic lemmas. After computing the number of new lemmas by CEFR level in our corpus, we can simply calculate how many meanings each lemma has at each of the CEFR levels. As expected, we found that the number of meanings per lemma increased as a function of the language level (see Figure 3). A learner starting to learn the vocabulary of a language will at first acquire just one meaning, the most basic one, per lemma. As vocabulary knowledge grows learners will not only acquire more lemmas, but they will also acquire more meanings per lemma. Going from the theoretical model to the corpus we used, we counted how many new meanings within that corpus would be acquired at each subsequent level. We found that indeed at the lowest levels the number of meanings is equal or almost equal to the number of meanings. By level A2 the average new lemma has about 1.67 new meanings and by level B1 and B2 the average new lemma comes with twice the amount of new meanings. As language users reach the C levels the average number of new meanings per lemma starts to reduce likely because rare or specialised words are less polysemous than more common words: for example, the word “parallelepiped” features in a dictionary in its unique geometric meaning, i.e. a prism whose faces are parallelograms.

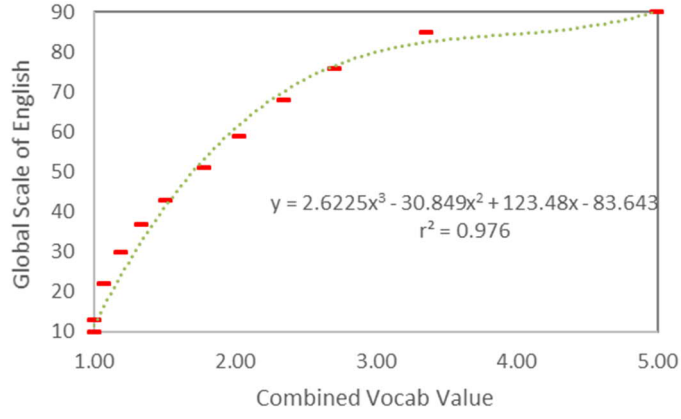
Figure 3: Vocabulary growth by level, by lemmas and meanings



After ordering the corpus by the combined values generated with formula 1 described in paragraph 4.3, we looked up which values corresponded to the hypothesised cumulative number of meanings by level. A mathematical model was then derived that

best described the relation between these values and the GSE values corresponding to these levels.

Figure 4: From Combined Vocab Value to GSE



Best fit was found with a third order polynomial function (see Figure 4):

$GSE = 2.6225\ C^3 - 30.849\ C^2 + 123.48\ C - 83.643$ (formula 2),

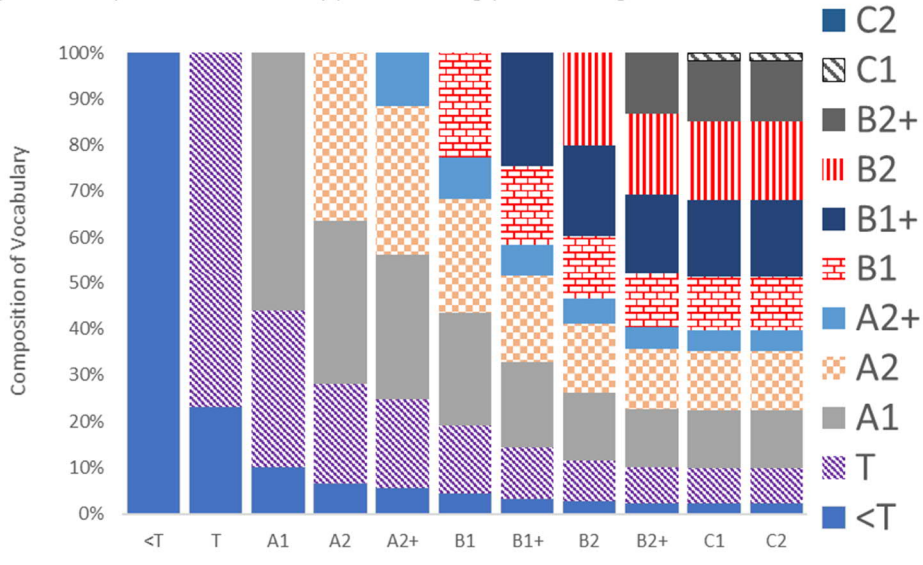
where "C" is the value combining the frequency data and the usefulness ratings. The explained variance for this function was very high ($r^2 = 0.976$). This function was then applied to compute the GSE values for all entries (word meanings) in the corpus.

4.5 Considerations

To understand the relevance of the frequency phenomenon, we evaluated the occurrence of word meanings at distinct levels of the CEFR and illustrated this in figure 5. Obviously, the typical entire vocabulary of a language user at the Pre-Tourist or "Start" level will consist for 100% out of word meanings at the Pre-Tourist level. Moving to the Tourist level, the language user will typically acquire about ten times as many word meanings, but the frequency of these new word meanings is on average lower than those acquired at the Pre-Tourist level. Therefore, although the number of word meanings at the Pre-Tourist level represents only 10% of the available word meanings for the Tourist level language user, because of their higher average frequency they will constitute about 25% of their total language use. Similarly, for example, the substantial amount of new word meanings acquired at the B2 level, represents almost 40% of all word meanings available to the language user at the B2 level, but because of their lower frequency these new acquisitions constitute only about 20% of their total language use. Moving to the C1 and C2 levels, it is useful to point out to learners and teachers that about 50% of the word meanings used at these levels are at B1 or below

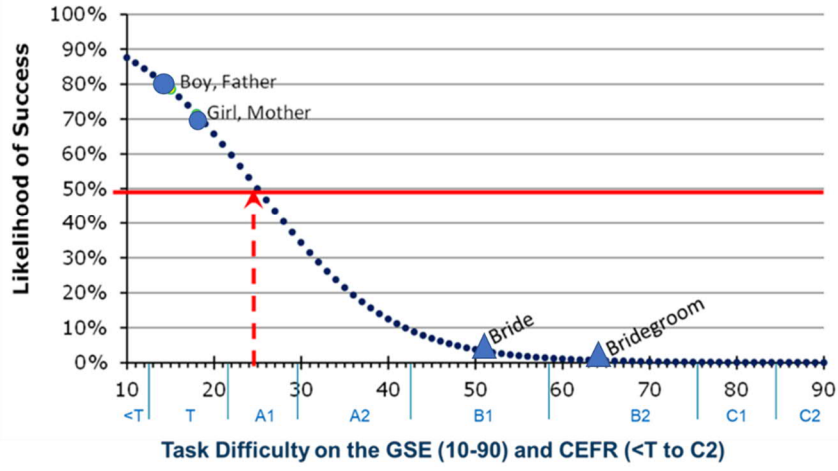
and that in fact the C1 level word meanings constitute only 2% of their word meaning usage. The proportion of C2 meanings at the C2 level is so small and their frequency so low, that they even do not show up in the graph in Figure 5. In fact, out of 100,000 meanings in their everyday average language use, speakers at C2 level will use only two C2 meanings.

Figure 5: Composition of vocabulary (word meanings) at increasing CEFR Levels



Finally, the development of the CEFR (North, 2000) and of the GSE (De Jong, Mayor, and Hayes, 2016) are both based on Item Response Theory (IRT), using the Rasch Model (Rasch, 1960/1980). The Rasch model is a family of psychometric models for estimating measurement properties from categorical data (such as the 1 to 5 ratings used in this study) as a function of a person's abilities, attitudes, or personality traits **and** the item or task difficulty. The GSE values assigned to a word meaning allow to estimate the likelihood that a language user with a particular level of ability will be successful in using that word meaning correctly. Figure 6 shows an example of a language user who is at 25 on the GSE scale, just above the lower boundary for A1, which is at 22. They have a likelihood of 79% of knowing the meaning of the words *boy* and *father* and of 71% of knowing the meaning of the words *girl* and *mother*. They would have started to obtain partial knowledge of the meaning of these words while at the Tourist level. They are not very likely (less than 5% chance) to know the meaning of the words *bride* and *bridegroom*. By reaching 51 on the GSE (B1+) they would have about a 50% chance of knowing meaning of the word *bride* and would probably only reach the 50% chance for the meaning of the word *bridegroom* by the time they are at 64 GSE, clearly above the lower boundary of B2 (59 GSE).

Figure 6: Likelihood of knowledge of word meaning by a learner at GSE level 25



5. Discussion

The present study reported on the statistical analyses carried out to combine frequency values and teacher ratings collected by Pearson to develop a vocabulary database aligned to the CEFR and the Global Scale of English. The database has been made available at pearson.com/english and is completely free to access. Of very large size, it provides information about more than 37,000 word meanings (corresponding to about 20,000 lemmas), 80,000 collocations and 7,000 phrases. Complementing the functional guidance found in the CEFR (Council of Europe, 2001) by providing lexical exponents for English, it is aimed to help language practitioners and researchers select level-appropriate vocabulary.

We believe our study innovatively contributes to the fields of language teaching and assessment for two main reasons. It provides a vocabulary framework for the English language which uses the word meaning as unit of count – rejecting the idea that an individual's vocabulary knowledge can be accurately interpreted using the lemma or word family as units of count, and therefore, taking into account polysemy and recognising that vocabulary usage cannot be detached from its contextual dimension. And it uses both a quantitative and qualitative approach by combining frequency data with judgements about the usefulness of words in order to produce a weighted measure to identify level-appropriate vocabulary. Since the CEFR levels and GSE values given to each word meaning are based on existing research into vocabulary size and since most of this research is concerned with comprehension (listening and reading) rather than production (speaking and writing), we provide an indication of the stage at which a particular word meaning can be expected to be part of the receptive vocabulary knowledge in terms of understanding word meanings by learners of English. Current insight estimates the productive knowledge to be about 50% of the receptive knowledge (Laufer and Goldstein, 2004; Shin, Chon and Kim, 2011).

The original approach of the Council of Europe provided extensive listing of lexical elements at the language levels A1 to B2 in the corresponding publications Breakthrough (Trim, 2009), Waystage (Van Ek, Alexander and Fitzpatrick, 1977; Van Ek and Trim, 1990), Threshold Level (Van Ek, 1975, Van Ek and Trim, 1990), Vantage (Van Ek and Trim, 1996) and it would be interesting to research how far the word lists in the four original Council of Europe specifications agree with the word meanings listed under the GSE values corresponding with CEFR levels A1 to B2.

A weakness of this study is its limitation in the size of the dataset. Brysbaert, Stevens, Mandera, and Keuleers (2016) estimated the median receptive vocabulary size of 20-year-old English first language speakers at 42,000 lemmas. Setting the much lower limit at 20,000 lemmas (see section 3) was necessary for this study to keep the work of categorising the lemmas, providing dictionary meanings and example usages and collecting the ratings for all word meanings manageable. It does, however, mean that from a theoretical perspective the number of word meanings for the C1 level and especially the C2 level is underrepresented. From a practical perspective we consider this a lesser concern: arguably language learners at these higher levels will be sufficiently equipped to decide for themselves on the word meanings they wish to acquire.

References

Anderson, R. & Freebody, P. (1979). *Vocabulary knowledge* (Tech. Rep. No. 135). Center for the Study of Reading, Urbana, University of Illinois

Baroni, M., Bernardini, S., Ferraresi A. & Zanchetta, E. (2009). *The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora*. In Language Resources and Evaluation 43 (3), 209-226

Benigno, V. & De Jong, J. (2017). *Developing the GSE Vocabulary*. Retrieved at https://prodengcom.s3.amazonaws.com/GSE-Vocab.pdf

Bogaards, P. & Laufer, B. (eds) (2004). *Vocabulary in a Second Language*. John Benjamins, Amsterdam

Browne, C. (2013). The New General Service List: Celebrating 60 years of vocabulary learning. In *The Language Teacher, 37*(4), 13–16

Brysbaert, M., Mandera, P., McCormick, S. & Keuleers, E. (In press). *Word prevalence norms for 62,000 English lemmas*. Retrieved at http://crr.urgent.be/papers/Word-_prevalence_norms_for_62K_English_lemmas_final.pdf (on 17/10/2018)

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

Council of Europe (2005). *Reference Level Descriptions for National and Regional Languages (RLD). Draft guide for the production of RLD: Version 2*. Language Policy Division DG IV - Council of Europe, Strasbourg. Retrieved at https://rm.coe.int/ 090000168077c574

Daller, H., Milton, J. & Treffers-Daller, J. (2007). Editors' introduction: Conventions, terminology and an overview of the book. In H. Daller, J. Milton, & J. Treffers-Daller (Eds), *Modelling and assessing vocabulary knowledge* (pp. 1-32). Cambridge University Press, Cambridge, UK

De Jong, J. H. A. L. (1989) Domeinbeschrijving en Toetsplan voor de Periodieke Peiling van Engels in het Basisonderwijs [Domain specification andtest description for the national

assessment of English as a foreign language in Dutch primary education]. Arnhem: CITO.

De Jong, J. H. A. L., Mayor, M. & Hayes, C, (2016). *Developing Global Scale of English Learning Objectives aligned to the Common European Framework*. Retrieved at https://prodengcom.s3.amazonaws.com/GSE-WhitePaper-Developing-LOs.pdf

De Jong, N. H., (2002). *Morphological families in the mental lexicon*. PhD Thesis, University of Nijmegen, Nijmegen. doi:10.17617/2.57697

Ellis, N. (2002). *Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition.* In Studies in Second Language Acquisition 24, 143–188

Granger, S. (2017). *Academic phraseology: A key ingredient in successful L2 academic literacy*. In Oslo Studies in English, Vol. 9, no.3, p. 9-27 (2017)

Gyllstad, H. (2007). *Testing English collocations* (Unpublished doctoral dissertation). Lund University, Lund

Hazenberg, S. & Hulstijn, J.H. (1996). *Defining a minimal receptive second language vocabulary for non-native university students: An empirical investigation.* In Applied Linguistics 17(2), 145-163

Laufer, B. & Nation, I.S.P. (1995). *Vocabulary size and use: Lexical richness in L2 written production*. In Applied Linguistics, 16, 307-322

Laufer, B.& Goldstein, Z. (2004). *Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness.* In Language Learning, 54 (3), 399-436

Meara, P. (2009). *Connected Words: Word Associations and Second Language Vocabulary Acquisition.* John Benjamins, Amsterdam

Milton, J. (2009). *Measuring second language vocabulary acquisition.* Multilingual Matters, Bristol

Milton, J. & Alexiou, T. (2009) Vocabulary Size and the Common European Framework of Reference for Languages. In Richards B., Daller M.H., Malvern D.D., Meara P., Milton J., Treffers-Daller J. (eds) *Vocabulary Studies in First and Second Language Acquisition.* Palgrave Macmillan, London

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* Rowley, MA, Newbury House

Nation, I.S.P. & Beglar D. (2007). *A vocabulary size test.* The Language Teacher, 31(7), 9-13

Norberg, C. & Nordlund, M. (2018) *A Corpus-based Study of Lexis in L2 English Textbooks.* In Journal of Language Teaching and Research, 9(3), 463-473

North, B. (2000). *The development of a common framework scale of language proficiency*. New York, Peter Lang

O'Loughlin, K. (2013). *Investigating lexical validity in the Pearson Test of English Academic Pearson*. Retrieved at https://pearsonpte.com/wp-content/uploads/2014/07/OLoughlin_ K_2014.pdf

Pearson (2000). *Longman Active Study Dictionary of English (LASDE).* Harlow, UK

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago, University of Chicago Press

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press, Cambridge, UK

Read, J. (2004). *Plumbing the depths: How should the construct of vocabulary knowledge be defined?* In P. Bogaards & B. Laufer (Eds.), Vocabulary in a second language: Selection, acquisition and testing (pp. 209-227). John Benjamins, Amsterdam

Roessingh, H & Elgie, S. (2009). *Early Language and Literacy Development Among Young English Language Learners: Preliminary Insights from a Longitudinal Study*. In TESL Canada Journal, 26(2), 24-45

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press, Cambridge, UK

Schmitt, N., Schmitt, D. & Clapham, C. (2001). *Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test*. In Language Testing, 18(1), 55–88

Schmitt, N. & McCarthy, M. (eds) (1997). *Vocabulary: Description, Acquisition, and Pedagogy*. Cambridge University Press, Cambridge, UK

Schmitt, N. & Zimmerman, C. B. (2002*). Derivative word forms: What do learners know?* In TESOL Quarterly, 36(2), 145–171

Shin, D., Chon, Y. V., & Kim H. (2011). *Receptive and productive vocabulary sizes of high school learners: what next for the basic word list?* In English Teaching, 66 (3), 127–152

Stubbs, M. (2002). *Words and phrases: corpus studies of lexical semantics.* Blackwell Publishing, Oxford

Trim, J. L. M. (2009). *Breakthrough: An objective at Level A1 of the Common European Framework of Reference for Languages, Learning, Teaching, Assessment (CEFR).* Unpublished. Retrieved at http://www.ealta.eu.org/documents/resources/Breakthrough_ specification.pdf

van Ek, Jan A. (1975): *The Threshold level in a European Unit/credit System for Modern Language Learning by Adults.* Council of Europe, Strasbourg

van Ek, J. A., Alexander, L.G. & Fitzpatrick, M. A. (1977). *Waystage English: An intermediary objective below Threshold Level in a European unit/credit system of modern language learning by adults.* Council of Europe, Strasbourg

van Ek, J. & Trim, J. L. M. (1990). *Threshold 1990*. Council of Europe, Strasbourg

van Ek, J. & Trim, J. L. M. (1990). *Waystage 1990.* Council of Europe, Strasbourg

van Ek, J. & Trim, J. L. M. (1996). *Vantage Level*. Strasbourg, Council of Europe. (Republished in 2000 as *Vantage*. Cambridge University Press, Cambridge, UK)

Widdowson, H. (2003). *Defining issues in English language teaching*. Oxford University Press, Oxford, UK

Wolter, B. (2009). *Meaning-last vocabulary acquisition and collocational Productivity.* In Fitzpatrick T. & Barfield A. (Ed.), Lexical Processing in Second Language Learners: Papers and Perspectives in Honour of Paul Meara (Second Language Acquisition), Multilingual Matters, Bristol 128-140

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge University Press, Cambridge, UK

Intellectual, insightful, inspiring – the ECML remembers Sauli Takala

Sarah Breslin, Susanna Slivensky and Michael Armstrong

ECML Secretariat

It is with great sadness that we at the ECML (The European Centre for Modern Languages) learned of Sauli's sudden death. In addition to the range and depth of his work at the University of Jyväskylä and around the world in the development of language education, especially in the fields of language testing and assessment, he also made an inestimable contribution to the work of the Council of Europe. Sauli was profoundly committed to the values of the Council of Europe and was involved in Council of Europe developments over several decades, most recently in the expert group for the development of the Companion Volume to the CEFR.

The ECML is indebted to Sauli: not only was he a member of the evaluation group which recommended that the ECML become a permanent institution of the Council of Europe but he also acted as programme consultant for the evaluation strand of the ECML 2008-11 Programme “Empowering language professionals”.

The work of the programme consultants is to accompany the different projects of the ECML's medium-term projects – giving advice and guidance, helping to solve any problems that arise and making sure that the projects combine practicality with academic rigour. In this work, Sauli was – as he was in all his activities – discreet, meticulous and thorough, prompting and stimulating rather than imposing a point of view. The project teams and the consultants felt privileged to have such an outstanding expert to help them to produce sound, useful, valuable projects.

Hanna Komorowska, Isabel Landsiedler, Marisa Cavalli and Frank Heyworth were Sauli's colleagues on the team of consultants and we were privileged to work with Sauli over 6 years. We treasured his kindness and quiet humour; he managed to combine scholarship and wisdom with informality and a quiet insistence on high standards; we all miss him as a colleague and friend.

Here are their personal tributes to Sauli:

From Hanna Komorowska:

I first met Sauli in 1971 at the International Seminar for Advanced Training for Curriculum Development and Innovation in Granna, Sweden, the event during which national teams of young researchers from many countries had a chance to learn from stars of the time – Benjamin Bloom and David (H.H.) Stern. Although the seminar attracted a large number of interesting personalities, it was impossible not to notice Sauli Takala from Finland. Sauli never spoke at length, but any time he decided to intervene, his comment would introduce a new aspect of the problem and his question would enliven or even redirect the course of discussion. I told him that he makes me think of Benya “the King” Krik, a hero of Isaac Babel’s "Odessa Stories" due to the way Krik was described by one of the characters, “Benya does not say much, but what he does say is savoury”. When, many years later, we started meeting as experts of the Council of Europe and consultants at the European Centre for Modern Languages I was happy to be in more regular contact with his admirable way of thinking and speaking - it was always an intellectual pleasure to listen to his creative ideas and precise comments. We miss them as we miss him.

From Isabel Landsiedler:

After having known Sauli Takala for his expertise for quite some time, I had the honour of working with him as a consultant for the Medium Term Programme of the European Centre for Modern Languages of the Council of Europe in Graz. In this role I was fascinated by his sharp arguments, his precise questions at exactly the right point in time and his useful contributions. He made incredibly sound comments, sometimes critical, but always constructive and right to the point. By interacting in this way he guided our meetings with his knowledge in his soft and friendly manner. His great power was to link concepts and find connecting points as he always saw the bigger picture and tried to look for more, some important issue behind the scenes. His energy to improve, innovate and change was so important for our meetings and for language learning in general. He was a very inspiring and powerful scientist that was so young and active in his mind. With his fine smile, his friendliness, his soft voice and his intelligent arguments he inspired people to care about language education, sometimes more than they would have done without him. I will remember his positive attitude, his spirit and his energy to innovate and improve forever, as he managed to open brains and hearts, which is a rare gift. Therefore, I am very grateful for having worked with him. Thank you, Sauli.

From Marisa Cavalli:

Je n'ai pas eu le plaisir de collaborer longuement avec Sauli Takala ni de connaitre ses travaux. Nous nous sommes professionnellement croisés. Les occasions de travail avec lui se comptent pour moi sur les doigts d'une main (trois, peut-être quatre). Mais elles ont suffi à me faire comprendre l'extraordinaire qualité humaine de Sauli, sa bonté exquise, son tempérament équilibré, intègre, accueillant et tolérant, la modération, mais la justesse aussi de son jugement. Je sais qu'il était une mine inépuisable de références scientifiques et bibliographiques : ses archives étaient redoutables de richesse et de précision. Elles faisaient le bonheur de ses collègues plus jeunes.

Je n'oublierai jamais l'accolade qu'il donnait, émouvante de chaleur humaine et d'empathie. Ni les derniers mots qu'il m'a dits la dernière fois que nous nous sommes rencontrés, quelques mois avant sa disparition : je les porte en moi comme un cadeau précieux. Il avait le don lui de voir et comprendre les autres, de les toucher à fond et de les encourager dans le chemin qu'ils ont entrepris.

From Frank Heyworth:

My first meeting with Sauli was at a meeting in Strasbourg in 1982, and I was struck by the way in which everyone listened carefully (not always the case in meetings in Strasbourg) to this rather unassuming-looking little man. Then, and over the years in countless meetings, I realised it was always because he was saying something constructive and insightful (again, not always the case). We worked more closely together in the years we were consultants together in Graz and I learned to appreciate his wry humour and the generosity with which he helped the projects he was involved with.

Anna von Zansen, ECML fellow for the project "A quality assurance matrix for CEFR use" remembers Sauli Takala:

My first encounter with Sauli's work was when I was training to become a language teacher in Finland. I then got to know him and his work more personally when I was working on the Digabi project at the Finnish Matriculation Examination Board. Sauli always had the time to answer my questions and he was interested in new projects. At conferences, Sauli often asked the best questions, ones that continued to inspire me long after the events themselves!

I can honestly say that Sauli's work has had a strong impact on my own professional development – in the classroom, in my research and on my fellowship at the ECML. I have been deeply affected by his death but am comforted by the conviction that his spirit and work will live on in the field of language education and testing. My condolences to everyone who had the honour to know Sauli.

CEFR-based language requirements in the European labour market

Cecilie Hamnes Carlsen
Western Norway University of Applied Sciences, Bergen, Norway

Bart Deygers
Centrum voor Taal & Onderwijs, KU Leuven, Leuven, Belgium

Beate Zeidler
telc gGmbH, Frankfurt, Germany

Dina Vîlcu
Babeș-Bolyai University, Cluj-Napoca, Romania

1. Introduction

Sauli Takala was a member of the Council of Europe Working Party involved in the elaboration of the *Common European Framework of Reference for Languages* (CEFR) in the early 1990s. He displayed a positive and optimistic view on the potential of the CEFR as a facilitator to international collaboration, yet, he emphasised the importance of well-informed and careful use for the CEFR to fulfil its purpose (Takala, 2007). Over recent decades, an increasing number of countries have set CEFR-based requirements in the language(s) of the host country for the purpose of residency, university admission, and access to employment (Deygers, Zeidler, Vilcu & Carlsen, 2018). And rather often, we see uninformed or even deliberate misuse of the CEFR-levels for such purposes (McNamara & Shohamy, 2008). Perhaps it was the urge to rectify unjust or uneducated applications of the CEFR that led Sauli Takala, together with Neus Figueras, to initiate and chair the European Association of Language Testing and Assessment (EALTA) CEFR Special Interest Group (SIG) in 2015.

Three of the authors of this paper are present or former chairs of the CEFR-SIG in EALTA's sister organisation, the Association of Language Testers in Europe (ALTE). Our concern aligns with that of Takala and Figueras: For the CEFR to be used in a way that truly promotes learning and collaboration, its users need to be well-informed about what the CEFR is, and what it is not. These are central questions in relation to both justice and validity.

In this paper, we will present a study conducted by the ALTE CEFR-SIG between 2016 and 2017. The purpose of the study was to obtain an overview of the ways in which the CEFR is used to set language requirements for entrance to the labour market across Europe. We were interested in knowing at what levels requirements were set (governmental, regional or local), in what professions it was most common to set specific CEFR-related language demands, and what levels of proficiency were normally required. The results of the study reveal substantial diversity in the practice of setting language demands in Europe. To supplement this initial stocktaking, the four authors, representing Norway, Belgium, Romania and Germany, describe CEFR-related language demands for professional purposes in their respective countries.

2. Use and misuse of language tests

Ever since Messick's seminal papers on validity in the 1980s and 1990s (Messick 1989, 1995), it has been widely accepted that language testers bear a responsibility not only for the internal qualities of tests, but also for the way test results are interpreted and used by society. Building on Messick, Shohamy (1990, 2017) argues that language test developers need a societal focus in addition to a psychometric one when striving for test quality and against the potential misuse of tests.

Insights related to the social impact of language tests have become guiding principles for many language testers, and have been embedded in the codes of ethics and codes of practice of three major language testing organisations; EALTA, ALTE and ILTA. Principle 9 of the International Language Testing Association's (ILTA) Code of Ethics, states that:

> Language testers shall regularly consider the potential effects, both short and long term on all stakeholders of their projects, reserving the right to withhold their professional services on the grounds of conscience. […] As professionals, language testers have the responsibility to evaluate the ethical consequences of the projects submitted to them.

In spite of these important efforts, language tests remain unregulated, and misuse can go unchecked and uncorrected (Spolsky, 2013). For that reason, it remains important to investigate the real-world use of language tests. CEFR-related language requirements for migrants are used in society for a range of purposes, some of which are defensible, while others are less so. The ethicality of setting language requirements for permanent residency and citizenship, for example, has been fundamentally critiqued (e.g., See McNamara, 2010; McNamara & Ryan, 2011; Pochon-Berger & Lenz, 2014). Language requirements for entrance to higher education or the labour market, on the other hand, have drawn less rebuke, presumably, because the importance of language proficiency in the target context can be demonstrated and operationalised. Most higher education systems are based on the majority language of the country, and therefore, in

order to be able to follow lectures, read curriculum, participate in discussions and write exam papers, foreign students need a certain level of proficiency in the target language. Similarly, for the labour market, it could very well be defensible that employers want to make sure that their employees have the necessary qualifications to carry out their job in a good way.

For entrance to higher education and the job market, the relevant question is therefore not whether setting language requirements is ethical or just in itself, but rather what level of proficiency is a necessary and sufficient level of language proficiency for the purpose. Nevertheless, also in these areas, uneducated, misguided or wilfully detrimental use of any language test should be rectified.

3. The CEFR and the labour market in Europe

In 2016 the ALTE CEFR-SIG initiated a study to investigate CEFR-based language requirements for the labour market across Europe. A year before, a similar study related to university admission language requirements had been carried out within the SIG (Deygers et al., 2018). That study concluded that most European universities have comparable admission policies, and most require international students to demonstrate B2 ability in the language of instruction. This chapter follows up on the university admission paper, and tackles two research questions:

RQ1: In which professions is it most common to set specific CEFR-related language demands?
RQ1: What levels of L2 proficiency are typically required for these professions?

Since the requested information is rather specific and the informants needed to be aware of the language regulations for professional purposes in their country, we chose the respondents via purposeful selection (Freeman, 2000). All respondents were ALTE representatives from one of the fifteen countries involved in this study. All respondents were professionally involved with language test development or research, and well informed on the CEFR and on the language testing policy in their country (median years of experience: 13, Min = 3, Max = 40). The countries involved in this study were Belgium, the Czech Republic, Denmark, Estonia, Finland, Germany, Italy, Lithuania, Norway, Poland, Portugal, Romania, Sweden, Switzerland, and the UK.

The electronic questionnaire consisted of three sections. In the first part, background information about each respondent was requested. The second section focused on the broad use of the CEFR in education and immigration, while the third and main section considered the use of the CEFR in the labour market at national and regional (depending on the political structure, this could be provinces, municipalities, and the like) level.

4. Summary and discussion of the questionnaire results

Most respondents claimed that the CEFR is not particularly well known in their country. Table 1 displays the responses of the 15 countries on a five point Likert scale (1=not well known at all, 5=very well known).

Table 1. How well known is the CEFR in your country? (N = 15).

Not well known at all	1
Not well known	8
Somewhat well known	5
Well known	1
Very well known	0

In all contexts surveyed the CEFR is used in adult L2 education, and in two contexts, the CEFR has inspired the L1 curriculum as well. In all fifteen contexts, CEFR-based tests are used in the immigration policy (i.e. requirements for permanent residency and/or citizenship). As has been observed before (e.g., McNamara & Shohamy, 2008), the levels required for this purpose vary substantially. In 11 countries, statistics were available on how many tests are administered. Combined, some 107,000 people every year take a CEFR-linked language test[1] for immigration purposes in one of the 15 countries covered by this study.

In sum, regarding these initial broad questions, there was some consistency across all fifteen contexts: even though the CEFR is generally not very well known, it is used to a considerable degree for setting language requirements for university entrance and immigration regulation.

Regarding access to the labour market, nine countries stipulate national requirements, and six do not. In four countries, regional authorities determine specific language-related employment criteria that differ from the national regulations. Typically, these regional requirements add to or specify national requirements. In one context (Italy) there are only regional requirements.

In ten countries, there are language requirements for certain professions, and in four of these countries, specific language tests are used. As Table 2 shows, these requirements most often focus on access to government administration or civil service (n = 6/10), teaching (n = 4/10), healthcare (n = 4/10), or maintenance and transportation jobs in the public sector (n = 2/10). In most contexts the required language level for civil servants and medical staff is proportionate to the level of seniority or responsibility. Nursing staff and low-level civil servants are typically required to

[1] The data are based on the informants' reported data, and the authors have not been able to check if the quality of the linking to the CEFR is appropriate as described in the Manual for relating examination to the CEFR (CoE, 2009).

demonstrate level B1 or B2, while C1 is the level required for high-ranking civil servants, doctors, and judges. In three cases, the language requirements were not stated in CEFR terms, but more impressionistic terminology (e.g., "satisfactory") was used. Overall, the requirements are not based on substantial empirical analyses. Eight out of ten informants claimed that the national requirements in their country were based on little or no empirical data or research.

The regional requirements are in line with the national requirements discussed above: They relate to healthcare, government administration and education, and are linked to specific CEFR levels. In one context (Norway) the municipal authority of the capital has implemented a B1-language requirement as a precondition for permanent contracts for all positions, when the applicants have not been schooled in Norway.

Table 2. Requirements at national level (n = 10 / 15).

	Total				
Government administration	6	B1-C1 (*n* = 2)	B2-C1 (*n* = 1)	B1-B2 (*n* = 1)	"satisfactory" (*n* = 2)
Teaching	4	B2 (*n* = 1)	B2-C1 (*n* = 1)	C2 (*n* = 1)	"satisfactory" (*n* = 1)
Healthcare	4	B1-C1 (*n* = 1)	B2-C1 (*n* = 2)	B2 (*n* = 1)	
Maintenance / transportation	2	A2 (*n* = 2)			

In all countries surveyed, individual employers have the right to stipulate language requirements for their staff. The survey data show that employers in the public sector quite frequently do so, but without a clear systematic approach and sometimes without reference to the CEFR. The results clearly indicate that, in most of the countries consulted, there are systematic and publicly available language requirements that regulate access to certain jobs in the public sector. The private sector, however, appears to be a different matter entirely. The data show quite consistently that the use and knowledge of the CEFR in the private sector is very limited, and that there is great variability in the language requirements private employers set.

It is striking to find that even though many professions do not use CEFR-related language requirements to determine access to the labour market, some sectors quite systematically do. This study shows that jobs in government administration, healthcare and education are uncharacteristically regulated, compared to other occupations. In most of the countries surveyed the language requirements in these professions become higher as the responsibility or status of the job increases. In all three professions in all ten contexts, the minimally required level is B1/B2, and the highest required level is C2. However, the requirements do not appear to result from extensive CEFR-knowledge or from empirical analyses. Moreover, access to the labour market was only

regulated by specific standardised tests in four countries, so it is impossible to gauge to what extent the level requirements are actually upheld in practice.

The data collected in this study show that the CEFR has perhaps not impacted the requirements for the labour market as much as it has the requirements for university admission. One reason for this could be that language testing for professional purposes covers substantially more circumstances, contexts and language proficiency levels than language testing for university admission. Another could be that not all employers see the value of setting systematic requirements. The CEFR is not used in all sectors or by all types of employers, but the general trend appears to be that public employers often rely on the CEFR to formulate requirements for civil servants, healthcare workers and teachers. The data offer no immediate reason to presume that employers from the private sector systematically use formal language requirements such as language tests or CEFR-based requirements. This is perhaps unsurprising, given that public accountability is generally a more pressing concern in public than in private sectors.

The general trends presented in this section offer only a glimpse of the overall picture of language requirements for the labour market in Europe. To supplement this overview, in the following section we will focus on specific cases: Norway, Belgium, Germany and Romania, and explore further the language requirements that grant access to the labour market in these four countries.

4.1 Norway: Responsibility for justice

In Norway, refugees, those granted asylum and family reunification have the right and obligation to follow 550 hours of Norwegian classes and 50 hours of knowledge of society (KOS) classes free of charge. Learners with low levels of literacy can acquire up to 3000 hours free of charge. The Norwegian classes are based on a curriculum for adult immigrants, which in turn is based on the CEFR. After the course, there is a compulsory test of Norwegian, *Norskprøven for voksne innvandrere* (hereafter Norskprøven) developed by Skills Norway (Kompetanse Norge) on assignment of the Ministry of Education and Research. Norskprøven is a high-stakes, standardised test measuring at levels Pre-A1, A1, A2, B1 and B2[2]. The test is also available as a proficiency test, and around 50% of the 20 000 candidates who take the test every year have not followed a state-provided language course first (Carlsen & Moe, 2016, 2017).

Norskprøven is a digital test that measures the four skills, reading, listening, writing and speaking in separate tests yielding independent scores. This flexible system allows users to set differentiated language requirements – for instance employers may require different levels of proficiency in different skills, or one may set requirements only in some skills and not in others.

An important question, then, is whether employers are familiar with the flexibility in the system as well as with the meaning of the CEFR-levels. A study carried out by Skills Norway investigated employers' knowledge and familiarisation

[2] A C1-level addition is being developed as this book is written.

with the CEFR and different tests of Norwegian at two points in time – in 2014, right after Norskprøven had been administered for the very first time and again three years later in 2017. 1 000 employers were phone interviewed at both instances. Surprisingly, the findings revealed no significant growth in familiarisation and knowledge between the two points (Haugsvær, 2018).

It is not feasible to present a complete overview of different language requirements in Norwegian society. There are few national or regional requirements, so by and large it is up to the different employers to set the requirement that they consider appropriate to their context. Instead of trying to give an overview, this section includes examples that show how test developers have taken their responsibility for justice seriously, striving to prevent misuse by informing employers about the content of the CEFR-levels and by opposing unjust and potentially detrimental use of test scores in society.

Two kinds of unjustifiable requirements in relation to the labour market have been witnessed: Firstly, that the proficiency level required is set at a higher level than is empirically defensible, and secondly, that written production is required for positions in which little actual writing is part of the job. Two concrete examples are given below.

In 2017, the municipality of Oslo proposed to introduce a B2 language requirement in all four skills for non-native speaking nursery school assistants. While it is reasonable and justifiable to demand a B2 level, also in written production, of pedagogical leaders in nursery school, a B2-requirement may be argued to be too high to be justifiable for assistants. Skills Norway responded to this proposition in different ways – by having a meeting with the municipality in order to inform the policy makers about the CEFR and the B2-level in particular, showing for example that only 3-4% of the 20 000 candidates who take Norskprøven each year reach this level in all four skills. As a direct result of the effort on the part of the test developers, the municipality decided to abandon the proposed B2-requirement and suggest a B1+ level instead. One may still argue that B1+ is too high, especially in written production, since the written tasks a nursery school assistant is likely to perform could be catered for with an A2-level in written production (write simple messages to parents and colleagues about routine matters, fill in forms).

Similarly, also in 2017, the county council of Buskerud proposed a B2-requirement in all four skills for taxi drivers. When arguing why they wanted to raise the language requirements, the council explained that they had considered a C1-requirement but had decided “only” to ask for B2. This in an obvious example of uninformed use of the CEFR and the potential detrimental consequences when decision makers lack the knowledge about the CEFR-levels and what they represent.

Again, through newspaper articles, radio interviews, and letters to the responsible policy makers test developers managed to convince the policy makers that a B2-level requirement was too high to be legally defendable for taxi drivers. As a response, the county council chose to reduce their requirement from B2 to B1 with explicit reference to the advice from the professional field.

These examples show that employers, also at municipality or county level, may set uninformed and potentially discriminating language requirements due to a lack of knowledge about the content of the CEFR-levels and the flexibility of the Norwegian test system. The examples also illustrate the importance of test developers taking responsibility for a just and informed use of the CEFR and of their language tests. Taking Messick and Shohamy seriously and striving to adhere to the codes of ethics and practice of EALTA, ALTE and ILTA means that test developers have a huge task also when it comes to informing about what their test measures, what the results mean and what would represent misuse of test results.

4.2 Belgium: Language tests and language politics

If language is political (Bourdieu, 1991), language policy can never be neutral (Shohamy, 2006). It is clear that language tests can be used as political tools, and that they can be used to selectively grant and deny certain people access to citizenship, housing, education, employment or other primary goods (Deygers, 2018; Deygers et al., 2018). In Belgium, language has probably been the most important source of political disputes, ever since the 1890s. Because of this, and because language proficiency in French and Dutch is demonstrably relevant in a bilingual context, high-ranking managers in the federal administration need to pass certain language requirements. This is one of the few professional contexts in Belgium with strict L2 requirements.

After more than a century of language-related political turmoil, the linguistic situation in Belgium has now reached a point of stability. Dutch, French, and German are the three official languages, and each of these languages has a legally demarcated territory. Brussels and a number of municipalities with special linguistic status are officially bilingual, but most of Belgium is either monolingually Dutch or French. In this situation of "territorial monolingualism" (Blommaert, 2011), the language of most of northern Belgium is Dutch while the south is mostly francophone, barring a small area where German is the official language. Each linguistic community is its own political entity, has its own government and parliament. Since linguistic communities and geographic regions do not always coincide, Belgium now has six different governments. One of those is the federal government which decides on certain matters (e.g., law, foreign policy, citizenship, etc.) but has no jurisdiction in others. The other governments (operating within the parliaments of Flemish government, the German-speaking community, the Francophone community, the Walloon region and the Brussels region) have the authority to decide on such matters as education, culture, and integration.

As a political entity, Belgium has formalised political linguistic compromises which have resulted in complex and detailed language laws. These laws detail who needs to know which language in a certain context and which types of proof are required. One of those contexts is the administration for the federal government which

employs some 70,000 employees (Blomme & Vervenne, 2017). The main language requirements/regulations apply to all of them, and the second language requirements to some of them.

Importantly, the main work language of every function in the federal administration has been determined by law. People applying for a job with a stated working language that differs from their language of tuition need to meet a language requirement. For some functions, a mandatory additional language has also been determined, and applicants will need to demonstrate language proficiency levels in this additional language as well. The situation is rather complicated and language legislation is a highly specialised matter, but in general terms it can be said that certain predetermined staff categories in the federal administration are required to prove a specific language proficiency level in Dutch (L1 speakers of French), French (L1 speakers of Dutch), or both (L1 is neither French nor Dutch).

The only way to prove that one has achieved the requirements is by passing the official language tests that are developed and administered by the federal government's selection and assessment agency, called *Federal Public Service Policy and Support* (*Federale Overheidsdienst Beleid en Ondersteuning*, or FOD BOSA). In some cases, the BOSA test can be waived when a commission accepts a specific language certificate from within the Economic European Area, or when a diploma offers sufficient proof of language proficiency. For many jobs in the public sector passing one of the BOSA tests is an essential requirement. Additionally, passing the right language test gives access to a monthly "language bonus" of up to €110. As a consequence these tests are high stakes and every year some 9500 people take them (SELOR, 2017).

In most cases, the language requirements and language tests are based on needs analyses: a linguistic profile of a certain position is analysed by consulting stakeholders and subject specialists. Type tasks are designed according to these profiles and then they are linked to the CEFR-levels. The test development and CEFR linking is done in-house, but expert teams of external test developers and researchers are routinely consulted for the purpose of quality control. In some cases, however, legal stipulations impact the test development process. And even though the entire test development team agrees with the external consultants that the test specifications are less than ideal, politicians and policy makers have the final say.

Perhaps the clearest example of this is the jurisprudence vocabulary test. The fact that it exists in its current form can only be understood by considering the Belgian political context in which language is a perennial political hot topic. In the 1960s, language politics were high on the agenda in Belgium. In 1962 the French-Dutch language border was formally decreed, and in 1966 the law on language use in civil service was passed. One paragraph in this law – added in 2003 – states that some federal civil servants need to pass a vocabulary test of legal terms. This jurisprudence vocabulary test, the paragraph stipulated, was to check whether civil servants understood legal terms and jargon equally well in both French and Dutch. Additionally, the test was to be administered orally in front of a jury. This jury reads forty legal terms in French or Dutch (whichever is the candidate's L1) in twenty minutes, and the

candidate is asked to translate it on the spot. The jury then consults on the quality and accuracy of the answers. The same law on language use in civil service also stipulated that staff members conducting staff appraisal talks need to prove oral B2 proficiency in French (if L1 is Dutch) or Dutch (if L1 is French). In this case, the law offered the test developers more flexibility, while also explicitly specifying the required CEFR level. Based on a needs analysis and an analysis of real-world texts, representative type tasks were identified, operationalised, validated and linked to the CEFR under the guidance of external academic consultants.

For years, this paragraph of the law was not implemented by lawmakers, but in 2017 it was revived for reasons of legal soundness. When it was finally addressed, the sensitivity of language politics meant that no political consensus could be found to update the law or the test requirements it stipulated. As a consequence, FOD BOSA was required to abide by the test specifications stated in the law in 2003 to develop an oral vocabulary test and an oral staff appraisal test. In the case of the former, this resulted in a jury-fronted vocabulary test, when a digital test could have served the purpose equally well, at a lower cost. In the latter the test developers were able to reconcile the law with the CEFR and with best-practice testing principles.

This brief introduction to Belgian professional language requirements offers one perspective on how political reality regulations can shape professional language requirements and language tests in one specific context. Perhaps most of all, this case shows how political pragmatism can override test development logic.

4.3 Germany: A migration hub

In 2016, 18.6 million people living in Germany - ca. 22% of the total population - had a migration background, i.e. they or at least one of their parents were not born German citizens. Many of these had obtained German nationality, while ca 10 million were still citizens of another state.

In politics immigration is sometimes viewed as problematic, but is also encouraged, as the influx of (in many cases well-educated) potential members of the workforce is welcomed by the business community. A number of organisations are involved in helping migrants to find work: first and foremost the Federal Office for Migration and Refugees (BAMF) and the Federal Employment Agency (BA), working on behalf of the Federal Ministry of Labour and Social Affairs (BMAS).

According to the EU regulated professions database, there are 149 regulated professions in Germany. For only some of these, linguistic requirements are specified. The regulated professions can be roughly classified as follows:

- Health care professions (doctors, veterinarians, general/geriatric nurses, paramedics, physiotherapists, health-care assistants, pharmacists, speech therapists etc)

- Pedagogical professions (school teachers, child-care workers, remedial teachers etc)
- Engineers including architects
- Lawyers
- Trades requiring a master certificate (master opticians, bakers, butchers, roofers, scaffolders, glaziers etc)
- Professions endowed with official functions (health inspectors, chimney sweeps, translators/interpreters with official function, etc)
- Other professions with enhanced responsibility (mountain guides, security guards, driving instructors, debt collectors, etc)

Regulations are either issued by the federal government or by the governments of the 16 states of Germany ('Länder'), and can be supplemented by the regional governorates ("Regierungsbezirke"), or by professional organisations. These regulations mainly concern the technical expertise required for each profession. As the 2017 Report on the Recognition Act states, there is a clear 'priority of checking qualifications over language skills' (2017 Report, 13). For the formal assessment of equivalence of a qualification to the corresponding German profession, language proficiency is not looked at. In some cases however, linguistic requirements have to be met before work can actually be taken up. Below, the linguistic demands in the most relevant regulated professions are briefly described.

- Doctors and pharmacists have to obtain a licence in order to be allowed to practise. The licence is issued by offices within the state governments. It can only be obtained if an applicant has the necessary proficiency in German, which proficiency this is can be decided by each office. However, in order to make regulations more uniform, the Conference of Ministries for Public Health issued a guideline in 2014 that specifies a level of at least B2 for general language plus C1 for the technical language of medicine. Psychotherapists, however, are expected to demonstrate specialist language skills at CEFR level C2. Which examination can be used as a proof of proficiency is then decided by the Medical Association of each state (2017 Report, 43).
- For general and geriatric nurses the target level is generally B2. As nurses are in high demand, B1 can be stipulated by individual German states in some cases (possibly with a specialisation in healthcare), as the Central Placement Office of the BA points out in its brochure.
- For school teachers, all German states specify CEFR level C2. For child-care workers, B2 is generally required, in some cases also C1.
- For trades requiring a master certificate, as well as for architects and engineers, there are no language stipulations.

Large numbers of migrants, however, take up work in unregulated professions. Available figures from 2008 show that at that time 14.4% of the employees with a migratory background worked in industrial production, followed by 11% in

'monitoring machines', 10% in the cleaning/waste disposal sector, and 8.1% in the food-service trade (Grünhage-Monetti 2010, 16). In these economic sectors there are no legal prerequisites, thus no legally stipulated language demands for taking up work. Employers are free to decide which language level they specify. In her study, Grünhage-Monetti conducted interviews with employees and employers, and found that migrants were employed even though their linguistic competence was felt to be lacking. Employers expressed the wish to receive support, e.g. by funding for language courses. In the meantime such projects have proliferated. Since 2017, work-oriented language courses are offered to participants who have successfully completed an integration course, and have reached level B1. Special modules will be provided for those who only reached A2 in the integration course.

In job advertisements, employers rarely seem to specify the required German language proficiency. A search in the database of the BA conducted on 4th April 2018 revealed that of 1,021,336 job offers, only 80,119 or roughly 8% required a knowledge of German, and of these only 410, or 0.04%, asked for any specific CEFR level.

Language requirements generally reflect a concern for the quality of the work that can be expected by the professionals in question, and the well-being of their clients. Here the CEFR is used to define a common standard. In non-regulated professions and trades, language requirements may exist, but they are rarely formulated in terms of CEFR levels.

4.4 Romania: Migration for specific professional purposes

According to the International Migration Report published by the United Nations in 2016 between 2000 and 2015 some countries have experienced rapid growth in the number of persons emigrating to other states. The Syrian Arab Republic (13.1 per cent growth per annum) is followed in these statistics by Romania (7.3 per cent per annum). In 2017 and 2018 nine persons left the country every hour, more than 3 million Romanians are currently working abroad. The economic migration, together with national factors, has intense and dramatic consequences for our labour market with the perspective of a human resource crisis in only a few years. Because of an imbalance between country regions and the drastically decreased birth-rate, hiring migrant workers has become a realistic prospect for Romanian employers. According to human resource specialists, in order to avoid a blockage in the labour market, Romania must employ a minimum of one million people from other countries in less than a decade. Currently, linguistic competence certification represents more a concern rather to those who want to leave the country than to foreigners who would like to come to work in Romania. The linguistic requirements for the persons who want to become Romanian citizens are not very precise either. According to the Romanian Citizenship Law, such a person "knows the Romanian language and has acquired basic notions of Romanian culture and civilisation, enabling him to integrate himself into the social life". To become a Romanian citizen, a foreign person will need to take an examination in

Romanian culture and civilisation in the Romanian language. However, no explicit language certification is required.

When people move to a different state for work or study, they try to enter into a developed country and once there they start to search for a job or new professional qualification or integrate into the education system. Linguistic competence is normally one of the key abilities they need to prove. In Romania the process of immigration happens differently, with the job market attracting exactly the people needed for certain jobs, either highly qualified, or barely qualified/unqualified. Whichever the case, the linguistic skills, whether significant, will be part of the initial deal. Thus, the necessity of linguistic skills will vary substantially from one case to the next, the professional qualification being of utmost importance instead.

According to the national laws regulating the labour market in Romania, private companies are free to employ people from outside the country in function of their needs. When it comes to language requirements, the law only mentions that Romanian is the official language of Romania, English is the most used foreign language and in multinational companies there is a demand for fluency in certain languages, like English, French and German. Numerous multi-national companies employ people from other countries for developing activities in their mother tongues in Romania and they do not require any certification of Romanian as a foreign language.

At the same time, the employers are concerned with the linguistic competence of their employees only when this is specifically needed for performing in the job. Whenever this happens, the employers do not seem to rely very much on tests in existence related to the CEFR or to a different framework. They prefer to have a job-related test, for any of the foreign languages they need their employees to master (the ones listed above, also some others – I know of the case of some companies in Cluj-Napoca contacting the foreign language centre of the Faculty of Letters for testing their employees in Norwegian, Swedish and Dutch - but not Romanian as L1) and for this they consult with the local testing institutions and contribute as major factors in needs analysis conducted by the test provider. Thus, if a test is taken by employees, the specific relation to the CEFR rather comes from the part of the test provider, already familiar with the framework and its advantages, rather than as a requirement from the stakeholders.

In the public sector foreign citizens can apply for any position, except for that of civil servant for which Romanian citizenship is required. In this case the legal provision that they need “to know spoken and written Romanian” refers to Romanian citizens belonging to different ethnic groups. However, the law does not mention in which way these persons need to demonstrate the required linguistic competence (e.g. through a certificate, an interview, etc.), nor does it specify the necessary level. According to the same law, civil servants need to be able to speak the language of a minority population, when this minority makes up for more than 20% of the total population in the region (Chapter X, Art. 108). In this case, also, no modality of demonstrating competence or level is specified.

As a general observation, the certificates of linguistic competence are admitted by employers rather due to the fame and reliability of the institution issuing them, so the validity of the certification depends on the commitment and fairness of the test provider in relation to the job done.

The job market for foreign citizens in Romania is growing fast. As a consequence, employers could become more aware of the linguistic competence their employees need. The CEFR could function as a firm point of reference for their different language needs and requirements. The examples of other countries show, however, that this is unlikely to happen in an educated, systematic or just way, if language testers and language testing specialists remain unconsulted.

References

Blommaert, J. (2011). The long language-ideological debate in Belgium. *Journal of Multicultural Discourses, 6*(3), 241–256

Blomme, P., & Vervenne, W. (2017, March 25). *Aantal federale ambtenaren duikt onder 70.000*. Retrieved May 1, 2018, from www.tijd.be.

Bourdieu, P. (1991). *Language and Symbolic Power*. (J. Thompson, Ed., G. Raymond & M. Adamson, Trans.) (7th ed. edition). Cambridge, Mass: Harvard University Press.

Bundesagentur für Arbeit, Zentrale Auslands- und Fachvermittlung, Ausländische Pflegekräfte für den deutschenArbeitsmarkt. Wie die ZAV Ihnen bei der Suche und Einstellung helfen kann. InformationenfürArbeitgeber, https://www3.arbeitsagentur.de/ web/wcm/ idc/groups/public/documents/ webdatei/mdaw/mjqw/~edisp/l6019022dstbai685070.pdf, accessed 4.4.2018

Bundesagentur für Arbeit, Working as an au pair for German families. Labour Market Permit, https://con.arbeitsagentur.de/prod/apok/ct/dam/download/documents/au-pair-in-germany-en_ba012998.pdf, accessed 4.4.2018

Bundesamt für Migration und Flüchtlinge, 2012, Berufsbezogene Deutschförderung. Das ESF-BAMF-Programm, http://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/Broschueren/broschuere-esf-bamf-programm.html?nn=1363754, accessed 4.4.2018

Bundesamt für Migration und Flüchtlinge, Konzept für ein Basismodul B2 im Rahmen der bundesweiten berufsbezogenen Deutschsprachförderung nach § 45a AufenthG, https://www.bamf.de/SharedDocs/Anlagen/DE/Downloads/Infothek/ESF/03_VordruckeAntraege/Deutschfoerderung45a/modulkonzept.pdf?blob=publicationFile, accessed 4.4.2018

Bundesamt für Migration und Flüchtlinge, Konzept für ein Basismodul C1 im Rahmen der bundesweiten berufsbezogenen Deutschsprachförderung nach § 45a AufenthG, https://www.bamf.de/SharedDocs/Anlagen/DE/Downloads/Infothek/ESF/03_VordruckeAntraege/Deutschfoerderung45a/konzept-basismodul-c1.pdf?blob=publicationFile, accessed 4.4.2018

Bundesministerium für Wirtschaft und Energie, Ausbildung und Beschäftigung von Flüchtlingen in der Altenpflege. Informationen für Arbeitgeber, https://www.bmwi.de/Redaktion/DE/Publikationen/Ausbildung-und-Beruf/ausbildung-und-beschaeftigung-von-fluechtlingen-in-der-altenpflege.pdf? blob=publication File&v=36, accessed 4.4.2018

Carlsen, C. & Moe, E. (2013) Assessing Norwegian. In Kunnan, A. (Ed.) *The Companion to Language Assessment. Volume IV*. Hoboken, New Yersey: Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla029,, accessed 01.04.2018

Carlsen, C. & Moe, E. (2016). Language testing as part of integration policy in Norway. *Kieli, koulutus ja yhteiskunta*, 7(6). https://www.kieliverkosto.fi/fi/article/language-testing-as-part-of-integration-policy-in-norway/, accessed 8.5.2018

Council of Europe (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, UK Cambridge University Press,

Deygers, B. (2017). Just testing. Applying theories of justice to high-stakes language tests. *ITL – International Journal of Applied Linguistics*, 168(2), 143–162.

Federal Ministry of Education and Research, 2017 Report on the Recognition Act, https://www.bmbf.de/pub/Bericht_zum_Anerkennungsgesetz_2017_eng.pdf, accessed 4.4.2018

Freeman, M. (2000) Knocking on doors: on constructing culture. *Qualitative Inquiry, 6*, 59–369.

Grünhage-Monetti, M. (2010), Sprachlicher Bedarf von Personen mit Deutsch als Zweitsprache in Betrieben, DIE, 12.07.2010, http://www.bamf.de/SharedDocs/Anlagen/ DE/Publikationen/Expertisen/expertise-sprachlicher-edarf.pdf?__blob=publicationFile, accessed 4.4.2018Haugsvær, K. 2018. *Rapport om arbeidsgiveres kjennskap til og bruk av norskrøver for voksne innvandrere* (unpublished report). Oslo, Norway: Kompetanse Norge.

ILTA Code of Ethics http://www.iltaonline.com/page/CodeofEthics, accessed 7.5.2018.

McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics, 18*(1), 89–95.

McNamara, T. (2010). The use of language tests in the service of policy: issues of validity. *Revue Francaise de Linguistique appliquée, 15*, 7-23.

McNamara, T. & Ryan, K. (2011) Fairness versus Justice in Language Testing: The Place of English Literacy in the Australian Citizenship Test. *Language Assessment Quarterly, 8*(2), 161-78.

Messick, S. (1989). Validity. In R. Linn (Ed.). *Educational Measurement*. New York, USA.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Pochon-Berger, E. & Lenz, P. (2014) *Language requirements and language testing for immigration and integration purposes.* Report of the Research Centre on Multilingualism. Universität Freiburg, Germany.

Romania (information sites, accessed 14.12.2018)
http://adevarul.ro/news/societate/noua-romani-pleaca-tara-ora-motivele-parasesc-romania-1_597766a85ab6550cb879ff5e/index.html
http://www.zf.ro/zf-24/romania-si-a-epuizat-resursele-interne-de-munca-singura-sansa-de-crestere-este-deschiderea-granitelor-pentru-a-aduce-un-milion-de-moldoveni-ucraineni-sarbi-italieni-spanioli-14709095
http://igi.mai.gov.ro/en/content/long-stay-visa-employment-secondment
http://incont.stirileprotv.ro/joburi-romania/tot-mai-multi-straini-vin-in-romania-sa-munceasca-unde-se-angajeaza-miile-de-chinezi-si-indieni-care-lucreaza.html
http://www.anofm.ro/files/Informare%20EN.pdf
http://www.avocatura.com/ll526-legea-188-r2-din-1999-privind-statutul-functionarilor-publici-actualizata.html
http://www.migrant.ro/citizenship-law

Shohamy, E. (2006). *Language Policy: Hidden agendas and new approaches*. London & New York: Routledge.

Shohamy. E. (1990). *The Power of Tests. A Critical Perspective on the Uses of Language Tests.* Longman. Harlow, England.

Shohamy. E. (2017) Critical Language Testing. In E. Shohamy et al (Eds.) *Language Testing and Assessment, Encyclopedia of Language and Education.* New York, USA: Springer.

SELOR (2017). *Rapportering 2016. Cijfers SELOR 01/01/2016 – 31/12/2016.* Brussel: SELOR.

Spolsky, B. (2013). The Influence of Ethics in Language Assessment. In A. Kunnan (Ed.). *The Companion to Language Assessment*. New York: John Wiley & Sons.

Takala, S. (2007) International co-operation in language education: before and after the CEFR. In Carlsen, C. & Moe, E. 2007. *A Human Touch to Language Testing*. Oslo, Norway: Novus forlag.

The EU single market regulated professions database, http://ec.europa.eu/growth/tools-databases/regprof/index.cfm?action=homepage, accessed 3.4.2018

United Nations Department of Economic and Social Affairs 2016. *International Migration Report 2015.* http://www.un.org/en/development/desa/population/migration/publications/migrationreport/ docs/MigrationReport2015_Highlights.pdf, accessed 14.12.2018.

Holistic peer analyses of National Tests in relation to the CEFR

Gudrun Erickson

University of Gothenburg, Sweden

1. Introduction

Sauli Takala took an active and positive interest in the studies reported on in the current text, regarding design and methodology as well as outcomes and implications. He often said that he appreciated tentative and exploratory studies like the ones focused on here, seeing them not as an alternative, but as an addition and complement to more standardized and traditional investigations. In particular, he emphasized the great value of collegial collaboration, which he considered a core aspect of the ethos of EALTA.

2. Background

Since the early 1980s, the Swedish national syllabuses for foreign languages have had a distinctly functional and communicative character, similar to what in the CEFR is referred to as an action oriented approach. Different revisions of the syllabuses (in 1994, 2000 and 2011) have made the relationship to the CEFR gradually more explicit, for example through the use of certain terminology and a slight shift of emphasis towards an even more competence and use-oriented view than before. However, the levels of proficiency required for the different steps in the Swedish system have not been fully aligned to the CEFR levels. Some textual analyses have been made (Börjesson, 2009; Hildén, 2008; Oscarson, 2002; 2015), as well as continuous observations in connection with the development and implementation of the national assessment system but, so far, no systematic, empirical studies of test results in relation to the Common European Framework, as described in the CoE Manual (Council of Europe, 2009) have been performed (for further information, see Erickson & Pakula, 2017).

The Swedish syllabuses define seven levels of foreign language proficiency, henceforward referred to as 'steps', common to all foreign languages[3]. As shown in Table 1, Step 2 in this system represents English as a foreign language (EFL) at the end of school year six (students aged 11-12) as well as so called Modern Languages, the

[3] Chinese is an exception, having a separate syllabus.

second foreign language (SFL) at the end of school year nine/compulsory school, usually French, German or Spanish[4]. Step 4 in the syllabus system represents English at the end of compulsory school (students 15-16 years of age). Modern Languages can be started either in Compulsory or in Upper Secondary School.

Table 1. The seven steps of foreign language proficiency in the Swedish national curricula (English and Modern languages); *Year* refers to Compulsory school, *Course* to Upper Secondary School.

Step	**English**	**Modern Languages**
1		Year 9; 'Students' Choice' (third foreign language; starts in Year 8; taken by < 2 %) Course 1
2	Year 6	Year 9; 'Language Choice' (starts in Year 6; started by c. 80 %, completed by c. 67 %) Course 2
3		Course 3
4	Year 9	Course 4
5	Course 5	Course 5
6	Course 6	Course 6
7	Course 7	Course 7

Each step is divided into different qualitative grade levels, before 2011 four levels, since 2011, six levels (A-F). Grading criteria, currently referred to as 'Knowledge requirements' (performance standards) define the lowest requirements for the different grade levels, with E as a minimal Pass (Swedish National Agency for Education[5]). In addition, there is an extensive system of national assessment materials and tests, the latter providing tasks, standards and a large number of commented benchmarks. All national tests are specified in relation to the syllabuses and developed according to a strict framework, which includes systematic collaboration with different groups of stakeholders (Erickson, 2017a; National assessment project website[6]). In the Swedish system, the national tests do not have the status of exams in the traditional sense but

[4] Modern languages can be started either in lower or upper secondary school; hence, all seven steps exist in upper secondary school, whereas step two is the final level in lower secondary school (for further information on SFLs in Sweden, see Bardel, Erickson and Österberg, 2019).
[5] https://www.skolverket.se/andra-sprak-other-languages/english-engelska
[6] https://nafs.gu.se/english/information/nafs_eng

are intended as advisory tools to be combined with teachers' continuous observations and assessments. However, their weight in the summative decision is not defined[7].

In preparation of the latest revision of the national language syllabuses (2011), there was a strong ambition at the national level to bring the Swedish system explicitly closer to the CEFR. This time, not only content standards were affected, but also performance standards, i.e. the levels of language proficiency defined in the two systems. Initiated by the University of Gothenburg, where the national tests were – and are – developed and funded by the Swedish National Agency for Education, a qualitative study was therefore designed aimed at investigating one of the national tests of English in relation to the CEFR. This, together with analyses of the national assessment materials for second foreign languages planned to follow, was meant as a tentative starting point for more systematic and empirical alignment studies. It is these studies, in particular the test of EFL at the end of compulsory school (step 4), that form the basis of the current text.

3. Study of the national test of English for school year nine

The compulsory national test of English for school year nine has, since its introduction in 1998, consisted of three parts, focusing on receptive competence (listening and reading comprehension), oral production and interaction, and written production and interaction. The oral and written parts (here referred to as Parts A and C) obviously require student constructed responses, whereas in the receptive part (B), there is an approximate 50/50 proportion between selected and constructed response formats. The oral part is a paired oral test, where students are asked first to talk on their own based on given prompts and then to interact with a peer, often to argue a case and to discuss certain suggested issues. The part focusing on writing has topics within different genres with a number of prompts and requires texts ranging from a minimum of approximately 100 to 400 words depending on the proficiency level being assessed[8]. Reception, as already mentioned, is the part that uses selected response for about half of the items, the rest being gap filling or short answer questions, only rarely requiring more than a few words. The test is accompanied by extensive scoring guides for teachers, who in the present system mark their own students' tests.[9] The guides include principles for rating as well as plenty of authentic examples and commented benchmarks for all parts of the test. Thus, these materials can also be regarded as serving an implicit function of rater training in a system where very much responsibility for assessment is placed on individual teachers.

[7] In 2018, a new regulation was introduced, stipulating that national test results shall be given 'special consideration' in teachers' grading decisions, however not further defined or quantified.
[8] Word limits introduced in 2013.
[9] Changes are underway, with anonymization of tests and other teachers than the students' own doing the marking.

3.1 Method

Twelve experts, all with profound professional experience of the CEFR, and of language testing and assessment, in twelve different European countries, were invited to participate in the study. The stated aim was to tentatively study the relationship between the national EFL test for step 4 in the Swedish system in relation to the CEFR, with regard to content and tasks as well as to standards. The study was intended to follow a basic, standardized scheme, however with ample opportunity for the participants to comment freely on any aspect they found relevant. With the aim of receiving independent judgements, the informants were told that they were part of an international group but were given no information about the other members.

3.2 Informants

Eleven of the invited informants gave a prompt and positive response to the invitation. At a somewhat later stage, an additional person working outside Europe declared interest in participating and was included as the 12th member of the group. Eventually, the group consisted of seven women and five men of varying ages, working in universities and schools, testing institutes or companies, and ministries. All informants had long experience of assessment (test development, policy work and/or research) in relation to the CEFR, and of various forms of international co-operation. In addition, some of them had been, or were at the time of the study, engaged in work for the Council of Europe and/or the European Commission.

3.3 Materials and instructions

The following materials were distributed to the participants:

- *Commentary letter*, providing a short description of the background to and aims of the study, the Swedish school system, incl. syllabuses and grading, and the rationale and function of the national tests;
- *Article* on the development of the Swedish national tests of Foreign Languages, published on the website of the national testing project at the University of Gothenburg (corresponding to Erickson, 2017a);
- *The Swedish national syllabus for foreign languages* (English translation), including the target level of EFL and the grading criteria for step 4;
- *The national test of English for school year 9, spring 2007* (all subtests);
- *Scoring guides*, including commented benchmarks, on paper and CD;
- *Response form*, comprising (1) CEFR scales considered clearly relevant to the different tasks in the test to be used in the overall comparison; (2) a table including the different parts/tasks of the test with instruction to note the estimated CEFR level of each task (not per item) regarding content as well as cut-scores/benchmarks. The response form also had plenty of space for comments on individual tasks as well as general observations and reflections.

Throughout the study, it was emphasized that what was intended was by no means an alternative to standard setting in relation to the CEFR, but a tentative study taking stock of the participants' individual knowledge and experience, and with the aim of informing the ongoing development work in and for language education within the Swedish system. One example of this message is the following, taken from the response form:

> *Please note that this is **not** intended to be **a standard setting exercise but a tentative, overall reflection** (at task and test, rather than item level), **based on your professional experience and expertise**.*

3.4 Results

The twelve informants delivered their analyses and comments in good order and within the time limit agreed. Self-evidently, the length and wealth of details of the reports varied, however with very good overall quality. Many aspects of interest, concerning test development, interpretations and applications of the CEFR, as well as of the Swedish system of testing and grading at large, were highlighted.

In general, the test as such, as well as the underlying principles and developmental processes, were considered positive. Aspects commented on were, for example, content in relation to target group, and variation and progression of difficulty. Several informants commented on the test as a whole providing ample opportunity for students at different proficiency levels to demonstrate their skills. The following quotation summarizes the view of several informants:

> *As said in the above, the tests are well made. There is sensible variation in terms of content, text types, test format. The texts seem to be related to the interests and cognitive level of the students for whom the test is intended.*

Others, however, strongly questioned the principle of the same test catering for the full range of proficiency within a cohort and also discussed the difficulty of relating a test of this kind to the CEFR.

Not surprisingly, a certain amount of variability of opinions was noticeable in the group of informants, both concerning the content of individual tasks and the variety of formats used in the test. For example, some considered a long listening comprehension task controversial, with regard to content (a story about a wolf attacking a child); others, however, classified the same task as exciting and engaging. Moreover, the proportion of constructed response items, as well as the choice between two different tasks for Writing[10], were discussed by single informants as potential threats to the validity and reliability of the results, the majority of the group, however, rather expressing positive opinions about these features of the test. In general, the agreement among the informants was somewhat higher concerning content and tasks than

[10] As from 2013, only one topic is given.

standards in relation to the CEFR. On the whole, however, the analyses clearly pointed in the same direction, which will be briefly exemplified in the following sections. What needs to be pointed out, though, is that the degree of detailing and specification varied considerably among the informants. Consequently, as pointed out initially, the study was indeed tentative, and no far-reaching conclusions should, or indeed can, be drawn on the basis of the results.

3.4.1 Part A – Focus: Oral interaction and production

The oral test was given a number of positive comments, exemplified by the following quotation:

> *A very well structured sub-test. Its construct and difficulty develops in a gradual way, allowing space for the least-able, the average and the best-able candidates to perform at their best. The test situation is authentic and life-like. One of the best features of the test is its paired/grouped format. The interlocutor intervenes only when she feels there is a break in the communication, or when there has not been enough speech performance elicited for reliable assessment.*

Quite contrary, however, one informant felt that the model as such did not *"offer much opportunity to demonstrate adaptation and variability"*.

Regarding the relation to the CEFR, the task/content level was considered to correspond to CEFR level B1, with a progression from a high A2 to a reasonably stable B2 in the three subsections of the test: individual production; argumentation followed by discussion.

Three authentic student conversations, intended to illustrate different levels of proficiency, were provided as benchmarks for rating. The informants' average rank ordering of these six students' performance levels proved to be the same as that provided in the scoring guides of the Swedish test, expressed by the informants as one pair representing an A2 and weak B1 performance, one illustrating stable B1s, and the final one considered representing the B2 level. (The issue of range will be further commented on in the concluding section of this text.)

3.4.2 Part B – Focus: Receptive skills

Part B of the test, focusing on receptive skills, was divided into two sections: part one for Reading comprehension, part two for Listening comprehension, both comprising varying content and formats. The Reading comprehension content level in the four tasks provided was classified as B1, with some progression between the tasks, the last one closing in on a low level B2. The Listening comprehension content level in two tasks was considered somewhat higher, the first one as a high B1, the second one as a low B2.

As previously mentioned, the tasks were perceived differently to some extent by the informants. A clear example is the second reading comprehension task, a gap

text in which twelve words had been rationally deleted. The comments from two informants illustrate the differences of views:

- *I liked this task a lot. The guidelines given are very detailed too.*

- *Seems to test inference, prediction and also writing. Because of its lack of focus I'm not in favour of this particular type of assessment.*

Another example is the listening comprehension task about the wolf mentioned previously, where the following two quotations exemplify the variability of opinions:

- *The format and its in-built variation are nice. The camping text narrates an exciting adventure and captures the interest. The items are well written. There is a nice mixture of selected and constructed answer items.*

- *Distressing text (should be avoided); too long; mixed formats; lack of standardisation can seriously jeopardise reliable assessment.*

The issue of content will be further commented on in the concluding section of the text.

In the Swedish system, the *performance standards* for Part B of the test, focusing on receptive skills, are based on aggregation of the results from the reading and listening comprehension sections (however, results presented in profiles, where each task is specified). Some of the informants found this very strange, whereas others considered it reasonable. Comments of the questioning kind often also included hesitation about one single test being used for all students, i.e. aimed at distinguishing between widely different levels of proficiency. Moreover, some informants abstained from commenting on the standards provided, instead advocating set percentage requirements for each level based on pre-testing data. In the light of this, a Pass level of approximately 35 % was considered too low, and even somewhat demoralizing. On the contrary, however, one informant, considered the tasks quite demanding and the standards a bit too high. The same informant also advocated qualitative judgements of difficulty in relation to standards, rather than what was characterized as a somewhat mechanical requirement of a set percentage for the different grade levels.

The informants who commented on the standards for Part B, focusing on receptive competence, gave the following average classifications:

Pass*: Low B1* (ranging from A2+ to B1)
Pass with distinction: *High B1* (ranging from B1/B1+ to B2)
Pass with special distinction: *B2* (ranging from B1+ to B2+/C1)

3.4.3 *Part C – Focus: Written production and interaction*

In Part C, two topics were provided for the students to choose between. Both tasks were quite open, one of them, however, providing more scaffolding than the other. Both genres of writing were included in the curriculum definition of written language, thereby justifying, or even requiring, alternative topics. The tradition, further underpinned by the national syllabus, was – and still is, to some extent – to favour texts of some length, in which students are asked to express themselves freely and extensively on different prompts, and to revise their texts before handing them in. Thus, having students write two different texts has not been considered quite doable within the given time span of 80 minutes. Also, in line with long-term praxis, no word limits were given (however, discussed as a possible addition)[11], and absolute task fulfilment was not a criterion for a pass, or higher grade. These conditions, especially the lack of word limits, made some of the respondents hesitant about the validity of the writing test, as summarized in the following quotation:

> *Having read the information about your syllabus and having carefully studied the benchmarks, I am more and more convinced that you would get much more reliable results and a much wider and more representative picture of the population's writing performance in the light of the syllabus if you asked candidates to do both tasks (with a length control of the performances) instead of offering the choice between them.*

Other informants, however, commented positively on the openness of the tasks and the challenges provided, especially in the second, more narrative subject.

The Part C task levels were considered to correspond roughly to CEFR level B1 for the first topic, some informants, however, considering even this task a bit more demanding (a high B1 or a low B2). The second topic was generally perceived as somewhat more demanding, by most informants estimated around a low B2 level.

Fourteen authentic student texts, seven for each topic, intended to illustrate different levels of proficiency, were provided as benchmarks for rating. The group's average rank ordering of these six students' performance levels was roughly similar to that provided in the scoring guides, however with noticeable variability, especially concerning a few samples. Some of the informants commented on what was characterized as a certain discrepancy between students' fluency and confidence, and their linguistic accuracy; two examples being the following:

> *Generally, I found the performances very positive and highly communicative. However, in several cases I was surprised at the lack of coherence and cohesion, as well as the great number of spelling and punctuation problems, among them several that hindered understanding.*

[11] As from 2013, both minimum and maximum word limits are provided, however expressed as recommendations rather than rules.

I think that Swedish students are a challenge to the CEFR, especially in Writing. They have fluency and use a wide range of vocab, structures – and make silly spelling mistakes and slip in terms of register. I suppose this has to do with the way they acquire the language and also with their cognitive development (they are very young), which makes it difficult to put them in slots for adults.

However, opinions varied considerably about this, one informant, who was very positive to the high demands on production in the test, pointing out that *"a higher level of accuracy seems to be expected than I associate with level B1"*.

Far from all informants commented on each individual benchmark, which is why only a comparison of the, at that time, three Swedish grade levels to the CEFR, not of individual texts, are presented here. The following average relationships were found:

Pass: *A2+/Low B1* (ranging from A1+ to B1+)
Pass with distinction: *B1+* (with a range from B1 to B2)
Pass with special distinction: *B2* (ranging from B1 to C1)

4. Small-scale study of national assessment materials for French, German and Spanish

Following the English study, and as part of the original plan to seek external comments on the national assessment materials in relation to the CEFR, a small-scale study of some of the materials aimed for second foreign languages in the Swedish school system was undertaken.

4.1 Second Foreign Languages in the Swedish school system

Studying a second foreign language (SFL) is optional in Swedish lower secondary school. This is a highly-disputed issue where, for long, opinions have been distinctly split among teachers as well as in political and policy oriented circles. However, a recent study (Erickson, Österberg and Bardel, 2018) indicates that teachers' opinions have changed considerably towards a much more positive attitude to making the study of a SFL mandatory in compulsory school. Furthermore, the Swedish National Agency for Education in a proposal to the Government in June 2018 suggested reforms pointing in the same direction (Skolverket, 2018[12]).

At present, around 80 per cent of all 12-year olds choose a SFL, usually French, German or Spanish (the latter by far the most frequent choice), but the dropout rate is considerable with only around two thirds of all students in each cohort leaving compulsory school having completed step 2 in the language syllabus[13].

[12] https://www.skolverket.se/om-oss/press/pressmeddelanden/pressmeddelanden/2018-06-18-forslag-for-att-fler-elever-ska-lasa-sprak
[13] For further information, see for example Bardel et al. (2019) and Tholin (2017).

4.2 National assessment materials for SFLs

Since SFL is an optional subject, the national assessment materials provided are not compulsory, but are offered to schools.).There are different kinds of materials, with formative as well as summative aims, the latter targeting the end of steps 2, 3 and 4 in tha language syllabuses. All three of these tests are used in upper secondary school, however, the one for step 2 being the only one also used in lower secondary school, at the end of school year nine.

The tests and scoring guides for the SFL tests are developed according to the same principles as the English tests, i.e. in a standardized, collaborative process including different rounds of pre-testing and analyses (Erickson, 2017a). Thus, what is offered to schools are subtests focusing on receptive, productive and interactive competences with the same type of tasks and formats that are used for English. A major difference, however, is that, according to a decision by the National Agency for Education, traditionally SFL teachers have been able – albeit not encouraged – to compose their own national tests by choosing from an electronic bank of standard-set subtests[14].

In connection with the standard setting rounds for the three languages, there is also what is referred to as 'horizontal validation' between the three languages undertaken. This means that the tests are compared pairwise by teachers having extensive knowledge in, and experiences of teaching two of the languages. In this comparison, a list of parameters is used, developed by an expert group and successively validated, focusing on different aspects of the tests regarding content as well as language. The aim of this procedure is to ensure, as far as possible, that the tests of the different languages, based on a common syllabus, are as similar as possible regarding level of complexity and difficulty.

4.3 Study of national assessment materials

Based on the positive experiences from the EFL study, it was decided that a small-scale SFL investigation with a similar design should be undertaken. In this case, however, only three European countries – one in the south, one in central Europe and one in the north, with one informant per language – were involved in analysing the national step 2 assessment materials of French, German and Spanish (taken by students at the end of Year 9/compulsory school or at the end of Course 2 in Upper Secondary School (see Table 1).

The study generated a large number of interesting comments and results. At a very general level, the CEFR-related outcome pointed in the direction of content as well as performance standards at an approximate, low A2 level. However, as with the English test, some of the tasks were seen as offering a clear opportunity to demonstrate

[14] This system is currently undergoing changes in the direction of unified tests comprising all three competences/subtests.

proficiency well above this level. In line with this, the performance standards for a Pass grade were most often classified as a very low A2.1 level, in some cases even lower, whereas the higher grade-level standards and benchmarks were often considered to be at B1 level (in a few cases even verging in on B2). Regarding the relationship between the materials for the three languages, no large differences were traced. However, it is important to emphasize that with as few informants as here – three (in one case four) per language – obviously, no strong conclusions can be drawn and no claims made on the basis of the results. Having said this, it can still be mentioned, as a point of potential interest, that there were some signs of the French materials being considered somewhat less demanding than the other two language materials, on the other hand, some German tasks in particular deemed quite advanced for an A2 level. This is obviously interesting, since the horizontal validations regularly undertaken as part of the test development and standard setting processes in Sweden have not pointed in the same direction. It is also an example of a phenomenon highlighted in the study that has been – and will be – further explored. One question that will have to be asked is obviously what can be related to the tests as such and what could perhaps also – or instead – be a result of different expectations for different languages/tests. As already said, though, the general and very clear picture was that of three reasonably parallel assessment materials all illustrating the A2 level. Finally, as in the English study, variability between the informants was quite clear, which is hardly surprising given the very different contexts they represented.

5. Summary and reflections

In the following, some of the observations made and opinions expressed by the informants in the two studies will be briefly commented on in relation to the actual tests as well as to the educational system in which they are developed and used. First of all, however, it should be stated that both studies worked very well for their tentative aims. Also, they strengthened what has been reasonably clear since long ago, both from a curricular and education-oriented point of view, namely that there is a strong resemblance between the Swedish language syllabuses and the CEFR. This applies in particular to the basic view of language expressed, with its clear action-orientation, thus also to assessment, regarding content as well as performance standards. This means that the different tasks as well as the Pass level for the Swedish national test of English at the end of compulsory school (step 4) is considered roughly equivalent to the B1.1 level in the CEFR, and the corresponding observations regarding the materials for French, German and Spanish step 2 indicates level A2.1.

Table 2 includes the conclusions drawn from the different rounds of textual analyses of the Swedish national syllabuses and the CEFR, published on the National Agency website[15]. The results of the peer analyses of the national tests focused upon in

[15] http://www.skolverket.se/download/18.6011fe501629fd150a28916/1536831518394/Kommentarmaterial_gymnasieskolan_engelska.pdf

the current text further support the model, which, however, does not build on the empirical validation of test data necessary to claim full alignment between the two documents.

Table 2. The estimated relationship between the seven steps of foreign language proficiency in the Swedish national curricula and the CEFR (based on textual analyses).

Step	Estimated CEFR level	English	Modern Languages
1	A1.2		Year 9; 'Students' Choice' (third foreign language; starts in Year 8; taken by < 2 %) Course 1
2	A2.1	Year 6	Year 9; 'Language Choice' (starts in Year 6; started by c. 80 %, completed by c. 67 %) Course 2
3	A2.2		Course 3
4	B1.1	Year 9	Course 4
5	B1.2	Course 5	Course 5
6	B.2.1	Course 6	Course 6
7	B.2.2	Course 7	Course 7

What needs to be emphasized, is that the estimations in Table 2 refer to a minimal Pass, hence a level intended to illustrate the bare minimum of what is required, not anything above that. Furthermore, it is important to bear in mind that the higher grade levels for each step in the Swedish language syllabus have not been systematically analysed in relation to the CEFR. Consequently, no empirically based answers can be given to questions about possible overlaps between grade levels above E for one step and higher steps/levels defined in the syllabus/CEFR.

The original aim of the two studies described in the current text was to get a tentative, external estimate of the relationship between the Swedish national tests and assessment materials and the CEFR. However, and in addition to this, the wealth of comments on various aspects of the materials provided plenty of interesting angles and interpretations that have been discussed from policy as well as test development points of view. As shown in the following, a rough division into three categories of comments can be made, namely comments on the context, the tasks and the standards.

5.1 Comments related to educational context

The comments related to the educational context, i.e. factors that emanate from decisions at the political level, often dealt with the range of the Swedish national tests, or more precisely with the fact that the materials in the Swedish system are aimed to tap widely differing proficiency levels. A certain hesitation was expressed by some

informants regarding the possibility of using the CEFR levels for a purpose like this. One of the effects of the Swedish system is that all students in the respective cohorts take the same test, with no adaptivity built in. As a consequence of this, one of the basic principles for developing the materials is that all students should be given as much possibility as at all possible to show what they can do with their language, not least by offering breadth and variation regarding the tasks provided. This is obviously a huge challenge, which requires collaboration with a large group of stakeholders, including teachers and students, and test development comprising an iterative process of piloting and large-scale pretesting. Included in this is collection, not only of traditional results from items and tasks but also of questionnaire data from students and teachers with the aim of letting their perceptions and suggestions feed into the development process and – as far as possible – influence the final products.

Another aspect of the Swedish system that generated a number of comments and reflections, is the fact that the national tests are not exams in the traditional sense but materials aimed to support teachers' own assessments and be combined with all other observations when teachers decide on individual students' final grades. Although currently undergoing distinct changes[16], the result of this is that, traditionally, the national assessment system has had explicit aims related to pedagogy and implementation as well as to fairness and equity (Erickson, 2017b; Gustafsson & Erickson, 2018). This has meant a certain emphasis on performance assessment and a somewhat lower degree of item-based testing and standardization than what many of the informants expected and were used to. However, it should be remembered that Sweden is by no means unique in having a system where teachers are deeply involved in large-scale assessment; this is, and has been, the situation in several other countries as well (East, 2015; Eckes et al., 2005); Spöttl et al, 2016), with varying experiences, partly due to traditions and context.

5.2 Comments on content and tasks

As exemplified in the text, opinions varied considerably concerning the content of some tasks. This reflects a common and important discussion on topics and texts used in assessment and testing, if what is deemed engaging content of different kinds is something positive or negative. Are students stimulated by exciting or even controversial texts or tasks, or is there a risk that their performances are affected negatively? Does 'neutral content' at all exist, and if this is claimed, neutral for whom and in what situations? Questions like these are of course not really possible to answer, since individuals are different and react differently to different kinds of input. However, there are broad guidelines to be found, for example in the Manual for Language Test Development and Examining, produced by ALTE on behalf of the Language Policy

[16] Following a national inquiry on the national testing system, a number of changes have been decided, e.g., about increased standardization. Also, a common framework at the system level has been developed, and is currently being implemented (Skolverket, 2017). Furthermore, digitalization of the system is underway and expected to be completed in 2022.

Division, Council of Europe (2011), where different examples are given of topics that may be seen as unsuitable for the intended target group: “war, death, politics, and religious beliefs, or other topics that may offend or distress some test takers” (p. 63). Especially the latter half of this quotation highlights the difficulty of defining what content is to be avoided in large-scale tests. In addition, there is also an educational dimension to consider, namely to what extent test developers are responsible for explicit or implicit messages conveyed, and for possible consequences at the individual and/or systemic level. Albeit complicated, there is however one fairly obvious way of handling, to some extent, the issue of suitability, namely to approach the real protagonists, i.e., the students themselves, and ask their opinions. This certainly does not solve the whole problem, but it gives very useful indications of possible reactions to different tasks and texts.

The example given in the current text of a listening comprehension task, whose content was characterized both as exciting and capturing interest and distressing (to be avoided) is a clear example of the problem of reaching agreement concerning type of suitable content. It may be of interest to know that there had been some initial doubts about this task in the project group, because of its length as well as its content. However, due to exceptionally positive opinions by a vast majority of students and teachers, as well as psychometrically very stable results in the pre-testing rounds, it was decided that the task should be included in the test. Post-test analyses of results and comments further strengthened this decision.

Finally, especially one type of comment about test format was interesting in relation, on the one hand, to tradition, on the other hand to empirical analyses of reliability and degree of acceptance. In the Swedish national tests, mixing selected and constructed response items in the same task is quite common, whereas in many other contexts it is seen as something less recommendable (however, usually not empirically supported). This was reflected in some of the comments given, both for the English test and the assessment materials for SFLs; whereas some informants found the variation of item types positive, others saw it as a potential threat to the quality of the tests. In this case, the development of tasks for the Swedish national tests has always relied on experiences from the use of the materials, where it has been very clear that tasks with mixed formats usually demonstrate very high reliability and also receive positive reactions from students and teachers. Furthermore, using both formats in the same task is also seen as a way of creating the breadth and variation deemed essential in tests aimed for very large and very mixed groups of students.

5.3 Comments on standards and benchmarks

As already mentioned, agreement between the informants tended to be higher for content than for performance standards in relation to the CEFR. What evidently was one of the most difficult aspects of comparing the two systems was the fact that the Swedish standards are meant to illustrate the bare minimum for its target level. Consequently, and fairly logically, the very lowest examples of a Swedish Pass for

Speaking and Writing were not always rated at the expected level (B1.1 for EFL/step 4, and A2.1 for SFLs/step 2), but one level below. As for the stronger examples/benchmarks in the Swedish system, i.e. those given grades above a Pass, there was no doubt that the expected CEFR target level was met, sometimes even higher than that. All this obviously also relates to the Swedish system of 'one test for all', which seemed quite new to most of the informants, in whose educational contexts praxis was rather 'one test per level'. This was even clearer in the item-based subtests focusing on receptive competence. Here, the expected levels were sometimes expressed in percentages of the maximum, with considerable variability between informants, where some informants required above 80 % correct answers for a Pass, whereas others thought that between 40 and 50 % (sometimes even lower) was enough to prove that the target level had been met. These observations hold true both for EFL and SFL. Also, certain differences in rater profiles could be noted, which is neither uncommon nor very surprising, especially in a tentative comparison of this kind, with no joint discussion preceding the estimations.

Finally, it was quite clear that, in spite of the clear action-orientation in the Swedish language syllabuses and the CEFR and consequently in the tests to be analysed, some of the comments on the productive and interactive parts of the tests indicated that accuracy was considered very important for the overall impression. The following examples may serve as an illustration:

> *Very fluent, good Vocabulary, good Interaction. Really a level B1, if the candidate had not made as many mistakes.*
> *If the student had made less mistakes, it would have been A2+ or B1.*

The same phenomenon has already been exemplified in an interesting way in a comment to the EFL Writing task, where one of the informants characterized the Swedish students as "a challenge to the CEFR" with their very uneven profiles regarding accuracy and fluency. Some comments on the tests of French, German and Spanish further emphasized this.

5.4 The European dimension

The current text deals with an attempt to tentatively relate the Swedish national language assessment materials to the Common European Framework of Reference for Languages, using peer assessment. In this, the EU initiated and funded European Survey on Language Competences (ESLC), undertaken in 2011, is of obvious interest (European Commission, 2012). In this study, 16 educational systems in 14 countries took part, each with their two most frequent foreign languages[17]. This meant English for all countries but the UK and most often French or German as the second language.

[17] Altogether, the five most frequent school languages in Europe were included in the study, namely English, French, German, Italian and Spanish.

Two countries took part in Spanish (France and Sweden), and only one in Italian (Malta). The survey, conducted at the end of compulsory school (ISCED 2), comprised tests of receptive competence and writing. As for reading and listening comprehension, only closed formats were used, mostly multiple choice, and writing was assessed through tasks with strict instructions. Speaking was not included in the ESLC, neither as production nor as interaction. All results were reported on the CEFR scale, which is of special interest in relation to the small-scale studies reported here.

The Swedish results differed very much for the two languages assessed. Whereas the English results were at the top as compared to the other participants, the Spanish results were very low, both compared to the other SFLs in the survey and in relation to the French results for Spanish. This has raised considerable attention in Sweden and has been discussed from a number of angles, not least including the shortage of certified teachers of Spanish in Sweden. Other aspects touched upon, have included, for instance, exposure to Spanish outside school, teaching traditions, and student motivation (see for example Riis & Francia, 2013). The fact that corresponding results for French and German in Sweden are not available is an obvious complication when it comes to interpreting the outcome, in particular focusing on the question whether the low Spanish results are subject specific or rather indicate a problem at the system level, with SFL not being mandatory in Swedish compulsory school. Attempts have been made to design a study of all three languages, using the ESLC instruments from 2011, but so far this has not succeeded. In relation to the small-scale studies discussed in the current text, the English results are not very surprising, which however can be said regarding Spanish, where there were no real indications that the tests were not considered suitable for the low A.2 level. However, repeating what has already been emphasized, it is essential to remember that with only three informants and with no joint training, variability is indeed to be expected and conclusions not really possible to draw.

Finally, it should be mentioned that there is interesting research available of studies where language levels in different countries have been successfully studied and compared using different tests, however designed, standard-set and reported based on the CEFR. One example of this is given by Hildén and others in the current volume, another can be found in Huhta (2016). This method obviously increases transparency and contributes to comparative analyses in an interesting and meaningful way. Indirectly, it could also be seen as a way of complementing other methods aiming to detect and establish relationships between national curricula and assessment instruments reflecting the CEFR.

6. Concluding remarks

The type of study described in the current text could be characterized as challenging, perhaps both in a positive and negative sense, due to a number of obvious differences at systemic, pedagogical and individual levels that make comparisons complicated. However, the aim was not to draw strong conclusions or to seek evidence for immediate

actions. Rather, the studies were designed to benefit from collegial reflections on issues of mutual interest, in particular regarding the relationship between different systems and materials, in this case the Swedish national syllabuses for foreign languages and the CEFR. Here, the results strengthened in a positive sense the impression of similarity between the documents and also indicated that the Swedish national tests correspond in a reasonable way to the intended target levels. Furthermore, reactions showed that several informants felt that the studies were interesting and useful also in their own contexts, which is obviously very positive. Hence, the outcome of the studies further emphasizes, and hopefully exemplifies, what is stated in the EALTA mission statement, namely that the purpose of the association is "to promote the understanding of theoretical principles of language testing and assessment, and the improvement and sharing of testing and assessment practices throughout Europe."

References

Bardel, C., Erickson, G. & Österberg, R. (2019). Learning, teaching and assessment of second foreign languages in Swedish lower secondary school – dilemmas and prospects. *Apples – Journal of Applied Language Studies 13(1), 7-26.*

Börjesson, L. (2009). *Jämförelse mellan skrivningarna i Gemensam europeisk referensram för språk (GERS) och kursplaner 2000 (7 steg) för språk* [A comparison between wordings in The Common European Framework of Reference for Languages (CEFR) and Language syllabuses 2000 (7 steps)]. Internal report for the Swedish National Agency for Education.

Council of Europe (2009). *Relating Language Examinations to the Common European Framework of Reference: Learning, Teaching, Assessment (CEFR). A Manual.* Retrieved in January 2019 from https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr

Council of Europe (2011). *The Manual for Language Test Development and Examining*, produced by ALTE on behalf of the Language Policy Division. Retrieved in January 2019 from https://www.coe.int/en/web/common-european-framework-reference-languages/ developing-tests-examining

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing*, *32*(1), 101–120. http://doi.org/10.1177/0265532214544393

Eckes, T., Ellis, M., Kalnberzina, V., Pižorn, K., Springer, C., Szollás, K., & Tsagari, C. (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.

Erickson, G. (2017a). *National Assessment of Foreign Languages in Sweden.* Retrieved in January 2019 from https://nafs.gu.se/information

Erickson, G. (2017b). Experiences with Standards and Criteria in Sweden. I Blömeke, S. & Gustafsson, J-E., *Standard Setting in Education. The Nordic Countries in an International Perspective* (pp 123-142). Cham, Switzerland: Springer International Publishing AG.

Erickson, G., Österberg, R., & Bardel C. (2018). Lärares synpunkter på ämnet Moderna språk – en rapport från projektet TAL [Teachers' opinions on the subject Modern languages – a report from the TAL project]. *Lingua, 2*, 8–12.

Erickson, G. & Pakula, H-M. (2017). Den gemensamma europeiska referensramen för språk: Lärande, undervisning, bedömning – ett nordiskt perspektiv [The Common European

Framework of Reference for Languages – a Nordic Perspective]. *Acta Didactica Norge 11*(3), 1-23. DOI: http://dx.doi.org/10.5617/adno.4789

European Commission. (2012). First European Survey on Language Competences. Final Report. Retrieved in Janury 2019 from https://www.researchgate.net/publication/262877352_First_European_Survey_on_Language_Competences_Final_Report

Gustafsson, J-E. & Erickson, G. (2018). Nationella prov i Sverige – tradition, utmaning, förändring [National tests in Sweden – tradition, challenge, change]. *Acta Didactica Norge 12*(4), 1-20. DOI: http://dx.doi.org/10.5617/adno.6434

Hildén, R. (2008). *Analys av svenska kursplaner i relation till den europeiska referensramen* [An analysis of Swedish syllabuses in relation to the Common European Framework of Reference]. Internal report for the Swedish National Agency for Education.

Hildén, R., Härmälä, M., Rautopuro, J. & Huhtanen, M. (2019). Finnish 9th graders' language skills: effects of learning environment and teaching on levels attained compared with other European countries. In Huhta, A., Figueras, N. & Erickson, G., *Developments in Language Education. A Memorial Volume in Honour of Sauli Takala* (pp. 113-130).

Huhta, A. (2016). Using the Common European Framework of Reference in the evaluation of educational achievement in foreign and second languages. In *Proceedings of the Second International Conference for Assessment & Evaluation: 'Learning Outcomes Assessment'*, Riyadh, Saudi Arabia, 1-3. December 2015. pp. 324-342. Available at http://ica.qiyas.sa/ downloads/Conference_Book.zip

National Agency for Education (2018). https://www.skolverket.se/andra-sprak-other-languages/english-engelska

Oscarson, M. (2002). *En Steg/Framework-jämförelse* [A Step/Framework comparison]. Internal report for the Swedish National Agency for Education.

Oscarson, M. (2015). Bedömning på systemnivå – En komparativ studie av stegsystemet i språk i den svenska skolan och språknivåer i Europarådets *Common European Framework of Reference for Languages (CEFR).* [Assessment at the system level – A comparative study of the 'step system' for languages in the Swedish school system and language levels in the Council of Europe *Common European Framework of Reference for Languages (CEFR).*] *Educare 2015*(2), 128-153.

Riis, U., & Francia, G. (2013). Lärare, elever och spanska som modernt språk: *Styrkor och svagheter – möjligheter och hot* [Teachers, students and Spanish as modern language: Strengths and weaknesses – opportunities and threats]. Uppsala University: Fortbildningsavdelningen för skolans internationalisering [Centre for Professional Development and Internationalisation in Schools].

Skolverket (2017). *Skolverkets systemramverk för nationella prov* [The Swedish National Agency for Education: System framework for national tests]. Retrieved in January 2019 from https://www.skolverket.se/publikationer?id=3890

Skolverket (2018). *Förslag för att fler elever ska läsa språk* [Proposal for more students to study languages]. Retrieved in January 2019 from

https://www.skolverket.se/om-oss/press/pressmeddelanden/pressmeddelanden/2018-06-18-forslag-for-att-fler-elever-ska-lasa-sprak

Spöttl, C., Kremmel, B., Holzknecht, F. & Alderson, J. C. (2016). Evaluating the achievements and challenges in reforming a national language exam: The reform team's perspective. *Papers in Language Testing and Assessment Vol. 5, Issue 1.*

Tholin, J. (2017). State control and governance of schooling and their effects on French, German, and Spanish learning in Swedish compulsory school, 1996–2011. *Scandinavian Journal of Educational Research.* 1-16. https://doi.org/10.1080/00313831.2017.1375004

Sauli Takala and his Archive – 'the love he bore to learning'

Elizabeth Guerin
University of Florence

For Sauli, in fondest memory

> *No man is an island, entire of itself; every man is a piece of the continent,*
> *a part of the main;*
> *if a clod be washed away by the sea, Europe is less,*
> *as well as if a promonotory were,*
> *as well as if a manor of thy friends or thine own were;*
> *any man's death diminishes me, because I am involved in mankind;*
> *and therefore never send to know for whom the bell tolls; it tolls for thee.*
>
> Meditation XVII, John Donne (1572-1631).

Any time one could meet and chat with Sauli was always an enriching experience to be savoured and cherised; one came away from Sauli with the feeling that he had not only listened to you and perhaps offered some suggestions if you had asked for advice, but aware also that Sauli would continue to reflect on the shared conversation and was likely to get back in contact with you when he had considered the interaction from other perspectives. Sauli's capacities for listening and reflection at length went well beyond the normal listening of the majority of people living in a world filled with stress and having 'no time' just now who say they 'will get back in touch' as soon as they can. Sauli was different! Sauli always had time to listen and help. Sauli exuded the peace and tranquility of his summer island retreat together with the reflective temperment of the quiet fisherman. Sauli transmitted, or better, tranferred these 'gifts' to those with whom he came into contact. That is one of the reasons why we miss him dearly!

The last time I had the privilege to meet and spend time talking with Sauli face-to-face was in Valencia at the 13th. EALTA Conference in 2016. This was a wonderful occasion which provided the opportunity to talk about and share the contents of his archive. This then stimulated me to reflect further on Sauli's involvement and research in different areas of language education, his recent work on the advancement of the CEFR Mediation descriptors - published posthomously, and propably one of his last official written contributions to language education and research on the same topic. This was also a moment of great insight into what an erudite scholar Sauli was, as well as what he deemed worthy of reading and referencing.

Sauli's archive consists of some 300 folders and 15,000 files, which is a remarkable number of texts even for a devoted lifelong scholar such as Sauli. The archive offers us – in case there were any need to do so! - a very good insight into the different disciplines and areas therein into which Sauli's acute mind delved in order to develop his – as well as our – understanding of the complex nature of language in its entirity. All of the above helps us to glean the enormous erudition of the scholar which lurked behind the humble and kindly smiling person we miss but who spurs us to further his varied research interests.

Sauli's research interests which were always indepth and marked by profound reflection spanned a broad area of interests as fitting of a scholar from a Humanistic background, and the readings in his reference archive bear witness to this. To borrow an expression, Sauli was a well-rounded Renaissance-type scholar. I dare to say this since his key reading and reference materials include papers from numerous disciplines such as philosophy, education, ethics, language acquisition methodology, gender, bilingual education, content-based instruction, content and language integrated learing (CLIL), language assessment – in all its aspects including classroom-based assement, language policy, teacher education, standards, technology, benchmarking, taxonomy, multilingual identities, multi- and plurilingualism, language competences, language for specific purposes, and, the list continues with etc. etc., because I am well aware that I have missed out on and glossed over some of his other related interests, as for example writing competence. In addition to the aforementioned areas of research, we also need to remember that Sauli studied these issues both *in* as well as *from* the perspective of different languages, because, for Sauli respect for the other – be it person or culture – was of paramount importance. Thus, though quite a remarakable achievement, it is not surprising to discover that Sauli was well able to communicate in quite a number of languages as if each of them were his mother-tongue. His linguistic expertise enabled him to broach different topics from multiple language perspectives. Indeed, it was not surprising to hear him switching languages in order to meet the needs of his interlocutor, show respect for other languages, and create interpersonal empathy. In his untold wisdom, Sauli realized the importance of the *indisoluble bond which unites culture with language as an indivisible whole.*

This language and culture ensemble is both highly relevant and extremely important in relation to Sauli the man and Sauli the scholar as is borne out in his research, as well as in his commitment and engagement in different contexts. Indeed, in order to contextualize (in the sense of Halliday) this concept and, so, gain a better insight into Sauli the individual and his scholarship, it is necessary to take a step back in history and consider some important and determining events which were to influence and shape Sauli's thinking.

In 1949, the Council of Europe (CoE) was set up with the objective of promoting human rights and pluralistic democracy. In 1954, the European Education and Culture Convention (EECC) was introduced. It was thanks to this Convention that, in the mid 1970s, Sauli became, as he used to say, 'engaged' with the CoE in the area of educational information and documentation (EUDISED), even though Finland

became a member of the Council only in 1989. This initial involvement was followed in the 1990s by what Sauli was wont to refer to as the beginning of his more ‘formal engagement’ in CoE language projects. His 'formal engagement' was marked by both intense and ongoing commitment and work with the CoE as his recent work – one of his last official contributions to language education and research – on the development and elaboration of the CEFR Mediation descriptors, shows.

Sauli's research interests, as we know, spanned many different areas of language and education. So what is it Sauli was always delving into at one point in time or another? His mercurial-like mind was constantly investigating, researching, and building knowledge in a systematic manner. By sifting through his huge ‘library’, it would appear that the way he went about his research was indepth and methodologic. By that is meant, that whatever the research topic uppermost in his mind in a given period of time, we can see how he builds up his reading materials and knowledge in that specific area. To give a more concrete example, we can trace his readings and research over different periods of time.

Sauli was very probably in a unique position as a researcher during his early research days at the Institute for Educational Research at Jyvaskyla University. The reason this conclusion is drawn is because, starting in the 1970s he was involved with foreign language curriculum development and evaluation in Finland. In 1974, together with Freihoff, Sauli produced 'A systematic description of language teaching objectives based on the specifications of language use situations' to be used in Higher Education, for the Finnish Ministry of Education. Given the period, this was an extremely forward-looking and revolutionary approach. General practice in teaching languages was based either on the pre-World War 'old key' grammar-translation approach or the approach developed in the 1940s based on a combination of structural linguistics and behavioural psychology, which with the entrance of America into the war, led to the Army Specialized Training Program (ASTP) which was based on the study and intensive use of drills which later became the Audio-Lingual method widely applied in language teaching from the 1960s. Indeed, those first scholars who advocated a approach to language learning based on what we can refer to as socially contextualized language in use situations, as Sauvignon (2018: 1) states, 'were met with skepticism, if not outright hostility'.

As we know from Sauli's research, we can place Sauli firmly within this group of forerunners whose thinking was harmoniously in tune with the contributions made to linguistics by the Prague School of functional linguistics (as opposed to Chomskian structural linguistics), and research by the sociolinguist Hymes (1972) on communicative competence.

Indeed, in the ‘70s, the CoE had begun a number of projects for young adults in the area of modern languages with the intention of developing language competences in an emerging Europe which was to promote mobility and social cohesion; language was the key to opening the door to integration (cf. Languages for Democracy and Social Cohesion (2014), and CEFR (2001: 1-9) for an introduction to important steps in the consolidation of CoE language policy over a sixty-year period).

As mentioned in passing in the previous page, this was the period (mid-1970s) when Sauli was 'engaged' with the CoE in the area of educational information and documentation (EUDISED). Hence, he was well aware of the issues involved in starting to build a culturally integrated Europe in states where WWII had seen bitter enemies destroy each other. Those enlightened people in the CoE were well aware that peace could only be stabilised through respect for each other and through cultural integration. This was no small task to tackle in a period of major emigration from the southern peripheral countries to the northern ones which offered work and the possibility to improve one's situation also through education. It is not by chance that, one of the most copious folders in Sauli's collection is that which contains publications related to the work and projects undertaken within the CoE; these publications start with those from that decade and continue thereafter.

As one, techically, on the extreme periphery of Europe, but in a country where all teacher education was university-based since 1973, Sauli was one who was well aware of the importance of language in building relationships and trust. Moreover, it is important to remember that Finland, given its historical background, was already a multilingual state with Finnish and Swedish embodied in constitutional law as the national languages of the new republic in 1917. Hence, it comes as no surprise then that Sauli had investigated the issue in depth in order to favour and facilitate the learning of further languages within the educational context. Already by 1980, Sauli had developed a general model of the language teaching system, which as Sauli himself states (1983:25), ‘is an adaptation of similar models proposed by Stern (1974) and Strevens (1977)’, as illustrated in the diagram in Fig. 1 below:

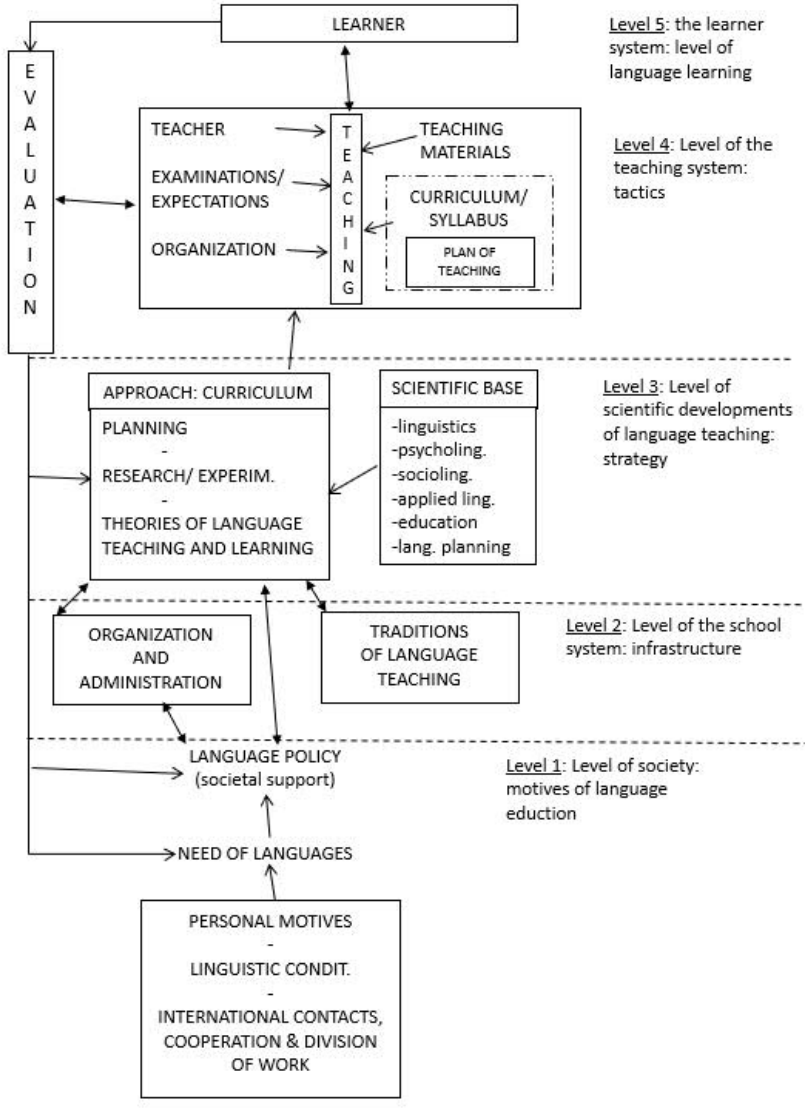


Figure 1. General Model of the Language Teaching System. Takala, 1980.

From Sauli's own account, finding himself in the 1980s for the first time at the University of Illinois at Urbana-Champaign was a challenge for him both from a teaching and a research perspective. Here he coordinated and also did research on the IEA International Study of Written Composition. Indeed, thanks to Sauli, we learn about interest in Italy, at hat time, in this area of research when he recalls the contribution of Pietro Lucisano at the European Centre for Education (CEDE) in Frascati in carrying forward this project work when times were tough and funding scarce (2011: 128). Indeed, the IEA folder is the other dense one in his collection.

Furthermore, it was in the 1980s while undertaking his Ph.D. research at Urbana-Champaign which he completed in 1984, that he was in close contact with Savignon (another researcher from Europe) in the area of Communicative Language Teaching. This was also the time language teaching (1980). In the field of bilingual education, another important development was taking place with Cummins (1979; 1981) who was formulating his theoretical proposition on Basic Interpersonal Communicative Skills (BICS) and Cognitive/Academic Language Proficiency (CALP) in L1 and L2 and their interdependent nature, as well as their implications for assessment. These were areas of research in which Sauli was also actively involved as his writings, as well as previous research in the Finnish context by Skutnabb-Kangas and Toukomaa (1976) demonstrate. The 1980s also saw Sauli in close contact with Trim and they both worked on functional linguistics and in the context of CoE Projects with people such as Piepho, Van Ek, Edelhoff, Trim etc. Sauli's writings during these two decades clearly acknowledge the influence of thinkers such as Bronowski, Husseral, Stuart Mills, Popper, Whitehead (Takala, 1982).

If we look at the period of the 1990s, when Sauli's more 'formal engagement' with the CoE was underway, Sauli's publications – in addition to those already mentioned – sshow his interest in action research (1994), bilingual education or Content and Language Integrated Learning (CLIL), and alignment. We can take a brief look at his research involvement in bilingual education, Content-Based Instruction (CBI), and Content and Language Integrated Learning (CLIL). As reported in Marsh, Oksman-Rinkinen and Takala (1996: 9-15), activities related to mainstream bilingual vocational education in Finland was not a new concept in the '90s. Already in 1991 CLIL was a reality especially in the universities of Jyväskylä and Vaasa. The university of Jyväskylä started Inset in CLIL in 1990 with a range of teacher development programmes aimed specifically at providing subject specialists with skills and knowledge to teach in a foreign language based on a theoretical approach using not only the Canadian immersion methods, but also, CLIL approaches used in South-east Asian countries such as Brunei, Hong Kong and Singapore. In 1992, Turku, following Inset training in CLIL for primary school teachers at the University of Jyväskylä, there was teaching in English at all primary levels from grades 1-6. Indeed, between 1991-1997, more than 700 subject and language teachers experienced Inset training in CLIL at the University of Jyväskylä, and the first significant research projects on CLIL in Finland started in 1996 (Marsh, Oksman-Rinkinen & Takala, 1996; Marsh, Nikula, Takala, Roviola & Koivisto,1999).

Alignment from, at least as far back as the 1980s, Sauli's interest in aligning test with the syllabus and curriculum is evident. Based on his archive, here Sauli drew on some 400 files divided amongst some 30 folders. The topics of these files range from policy briefs, to the theory of systemic reform, to the alignment of curriculum standards and assessment, and, later, to linking examinations to the CEFR (2004; 2009). In the timespan prior to and following the seminar organized in Finland in 2002, while Sauli was working with other CoE colleagues (Figueras, North, Verhelst, Kaftandjieva, etc.) on the development of the Manual for Relating Examinations to the CEFR which

was published in a preliminary version in 2004, as well as the Related Supplement, with the final version of the Manual published in 2009. He was also involved in the Council of Europe´s work on modern languages, as well as the EU-funded DIALANG project coordinated by my home department, Center for Applied Language Studies, University of Jyväskylä, coordinated during the first phase (autumn 1996 – November 1999). Moreover, Sauli became President of the European Association for Language Testing and Assessment (EALTA) for the period 2007-2010. Unfortunately, it was during his term in office, in 2009, that Sauli's close colleague and like-minded spirit, Dr. Felly Kaftandjieva, passed away unexpectedly; this was a great loss for all who knew this kind, respectful and gentle soul. In her memory, in 2010, the Felianka Kaftandjieva Memorial Lecture was established.

There was at least an occasion, once a year, when many of us had the opportunity to enjoy the acumen of Sauli's mind: the annual EALTA Conference. This event was a part of the life-journey Sauli shared with many people. Indeed, it was in May 2004, at the first EALTA Conference in Kranjska Gora, about which I learned thanks to Neus Figueras – with whom I was collaborating on the EU CEFTrain Project – that I had the opportunity to met Sauli for the first time. From thereon in, I was a regular and looked forward to meeting and exchanging ideas with Sauli during the EALTA conference each year. He opened up the real world of assessment to me, and I had an awful lot to learn! He was elected the second President of EALTA in 2010, I was privileged to participate in EALTA's first Summer School on '*Good Practice in Language Testing and Assessment: An Educational Perspective' which* took place at the Norwegian Study Centre, at the University of York (UK). This initiative enabled me to understand Sauli's keen mind and learn about the important details and fairness in the field of testing. As a teacher, Sauli explanations of deep topics in apparently simple terms and his 'plain talk' enabled neophytes to grasp an understanding of complex ideas and topics in a friendly and non-challenging way. No question asked by participants was too trivial to be addressed seriously in simple terms, and Sauli was always ready to guide one in the right direction to find the solution to the problem that was bothering one, or to the readings that would inform one on the issue at hand. As Sauli (1983: 33) states:

> Educators should not underestimate the positive contributions of evaluation, as they should not underestimate the possible negative washback effect of evaluation that is not congruent with teaching objectives and the teaching itself.

And Sauli always did 'practice what he preached'. For Sauli, the details were always important …, they were 'what made the difference'!

The part of the road on our journey's way shared with Sauli was wonderful. In Valencia, I was happy to share with him the QUAMMALOT Project that we had submitted to the EU for funding and which aimed at developing a qualification for teachers of migrant minors to facilitate the integration of unaccompanied migrant minors in schools with colleagues from Spain, Greece, and Denmark, as well as Italy.

He was enthusiastic about the idea because of his profound sense of respect for others and otherness, as his deep interest in mediation testifies. Though, unfortunately, I never got the chance to share with Sauli the approval of the project, somehow I feel he knows and is smiling at it somewhere.

All of the above helps us to glean the enormous erudition of the scholar which lurked behind the humble and kindly person we miss but who spurs us to further his varied research interests. As we find in a few of the EALTA website tributes for Sauli:

> He had an open mind, a warm heart and a winning personality.
> Sauli, thanks for having been who you were, we will miss you always.
> We have lost a friend but we will not forget you.
> (https://ealtasaulitakala.wordpress.com/2017/02/16/first-blog-post/tributes)

Having reflected for a little on the kind and affectionate person we were fortunate to meet on our life-journey and share some treasured moments with, remembering his numerous contributions, his immense culture and scholarship, as well as his sincerity and simplicity, it is fitting to draw this memory of Sauli to a close with some thoughts of his:

> Language teaching is therefore not only the activity of individual teachers; it is a system of many activities. To understand it as a system, we must realize its boundaries, its central purposes, and its level in a larger context. We must be aware of its various subsystems and their interrelationships. For all this we need models to describe and work out the practical consequences of different approaches.
>
> (Takala, 1983: 25).

References

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Strasbourg: CoE.

Council of Europe. (2014). *Languages for Democracy and Social Cohesion: Diversity, equity and quality. Sixty years of European cooperation*. Strasbourg: CoE.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, No. 19, 121-129.

Cummins, J. (1981). Empirical and Theoretical Underpinnings of Bilingual Education. *Journal of Education, 163*(1) 16-29.

Figueras, N., North, B., Takala, S. Van Avermaet, P. & Verhelst, N. (2009). *Manual for Relating Examinations to the "Common European framework of Reference for Languages"*. Preliminary Version. Strasbourg: CoE.

Freihoff, R., Takala, S. (1974). *A systematic description of language teaching objectives based on the specification of language use situations*. Jyväskylä: Language Centre, University of Jyväskylä.

Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–93). Harmondsworth, England: Penguin Books.

Marsh, D., Oksman-Rinkinen, P. & Takala, S. (1996). (Eds.) *Mainstream Bilingual Education in the Finnish Vocational Sector.* Introduction – Education in a foreign language: an old concept. National Education Board.

Marsh, D., Nikula, T., Takala, S., Roviola, U. & Koivisto, T. (1999). *Language teacher training and bilingual education in Finland*. Acc. 15/05/2006. (http://web.fu-berlin.de/elc)

Savignon, S. J. (1972). *Communicative competence: An experiment in foreign language teaching*. Philadelphia, PA: The Center for Curriculum Development.

Sparrow, J. Keynes, G. (eds.) (1923). *Donne's Devotions Upon Emergent Occasions* by John Donne. Cambridge: Cambridge University Press.

Skutnabb-Kangas, T. & Toukomaa, P. (1976). *Teaching migrant children's mother tongue and learning the language of the host country in the context of the sociocultural situation of the migrant family.* Helsinki: The Finnish National Commission for UNESCO.

Strevens, P. (1977). *New Orientations in the Teaching of English*. London: Oxford University Press.

Stern, H. H. (1974). Directions in language teaching theory and research. In Qvistgaard, J.

Schwartz, H. & Spang-Hansen, H. (eds.) *AILA Third Congress Proceedings', Vol. III: Applied Linguistics, Problems and Solutions*. Heidelberg: Julius Groos Verlag.

Takala, S. (1980). New Orientations in Foreign Language Syllabus Construction and Language Planning: A Case Study of Finland. Institute for Educational Research, University of Jyväskylä, Bulletin No. 155. (Also in ERIC ED 218 925).

Takala, S. (1982). The Need for Theoretical Advance in Education and in Language Education. *The Rackham Journal of the Arts and Humanities*, Vol. 2 (3) 105ff.

Takala, S. (1983). Contextual Considerations in Communicative Language Teaching, In S.J.

Savignon & M.S. Berns (Eds.). *Communicative Language Teaching: Where Are We Going?* Illinois Univ., Urbana. Language Learning Lab.

Takala, S., Verhelst, N., Kaftandjieva, F. & Banerjee, J. (2004, rev. 2009). *Reference Supplement to the Manual for Relating Language examinations to the CEFR*.

Takala, S. (2011). The International Study of Writing. In C. Papanastasiou, T. Plomp & E.C. Papanastasiou (Eds.). (2011). IEA 1958-2008: *50 Years of Experiences and Memories* (115-136). Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

What it means to be at a CEFR level

Or why my Mojito is not your Mojito –
on the significance of sharing Mojito recipes

Claudia Harsch

University of Bremen

1. Introduction

The Common European Framework of Reference (CEFR) is at its core a language policy instrument of the Council of Europe. In its nine chapters, it sets out to describe relevant aspects of language learning, teaching and assessment (in this order). One aspect that had a tremendous impact at least in Europe is the CEFR's conceptualisation of learner language as authentic language. This conceptualisation draws the focus on what learners can already do with a foreign or second language and how well they can do it. The development of the European Language Portfolio is but one illustration of the shift towards greater learner autonomy that was brought about by the CEFR (e.g. Little, 2005). The focus on positively describing learner language stipulated a rethinking in many European educational systems on how learner language and learner achievement is conceptualised and operationalised in curricula, classrooms and exams. The CEFR has informed educational reforms across Europe (Broek & van den Ende, 2013), for instance in Austria[18], where the Matura for foreign languages was reformed between 2005 and 2009; in Finland, where the Finnish School Scale was related to the CEFR in the early 2000s (Hilden & Takala, 2007); in Hungary[19], where school-leaving exams for English as a foreign language were reformed between 1998 and 2002); or in Germany, where Educational Standards for the foreign languages based on the CEFR were developed between 2003 and 2012[20], and test instruments aimed at monitoring educational achievement were developed and aligned to the CEFR (cf. Rupp, Vock, Harsch, Köller, 2008; Harsch, Pant, Köller, 2010).

This chapter is dedicated first to what the CEFR as a framework of reference for language learning, teaching and assessment can do for language educators, and where its limits are. Regarding the CEFR's limitations, I will then take the realm of language testing under closer scrutiny, as this is an area where the CEFR had an immense and critically disputed impact (cf. e.g. Alderson, 2007; Weir, 2005) that reaches far beyond

[18] For more details (in German), see https://www.uibk.ac.at/srp, accessed 26.03.2018.
[19] For project details and outcomes, see http://www.lancaster.ac.uk/fass/projects/examreform/Pages/Projects.html, accessed 26.03.2018.
[20] For more details (in German), see https://www.kmk.org/themen/qualitaetssicherung-in-schulen/bildungsstandards.html, accessed 26.03.2018.

Europe. For example, all major internationally operating exam providers have aligned their tests to the CEFR. Alignment to the CEFR, however, does not mean that the tests measure the same or that the tests result in equivalent classifications of learners with regard to the CEFR levels. I will examine potential reasons for the hotly debated issue that different tests may yield different results despite having been aligned to the CEFR, and I will give an outlook of how we can deal with these discrepancies.

2. The need for localising the CEFR

The CEFR as a common framework can serve a variety of functions. Amongst others, it can inform learning and teaching goals; provide a basis for curriculum development and educational standards; facilitate constructive alignment of learning, teaching and assessment; be used as a starting point for defining assessment constructs; or inform learner- and teacher-oriented assessment.

With regard to the proficiency framework offered in the CEFR's chapters 4 and 5 and the recently published CEFR Companion (Council of Europe, 2018), the scales and descriptors presented there can help to specify learning and teaching aims, define constructs for communicative language assessment, and inform teacher-/learner-oriented assessment. This is facilitated by the CEFR taking a multifaceted, hierarchical view on proficiency, which ranges from a global, overall perspective to ever more detailed facets of language proficiency. For each of these hierarchical perspectives, the CEFR offers illustrative scales with descriptors on six ascending proficiency levels. The proficiency descriptors offer a common meta-language to communicate curricula content, expected learning/teaching outcomes, educational standards and assessment criteria, thus enhancing communication among stakeholders. The framework initially had raised hopes that language educators may come to a shared interpretation of the CEFR levels, assuming that "my B1 is your B1". I remember one conference presentation entitled "Is my Mojito your Mojito?" (Avermaet, 2004), implying that while there are local variations of Mojitos, we all recognise a Mojito, just as we then hoped that we all would recognise "a B1 performance". The CEFR, however, cannot satisfy this hope for a variety of reasons that I will outline in this chapter. What we realistically can hope for is to share our individual interpretations of "what my B1 is like", i.e. how we, in our local contexts, interpret the CEFR levels. With reference to the Mojito metaphor, we are now at a point of acknowledging that we have to share our individual recipes for our local variations of Mojitos in order to make transparent what our interpretations of a Mojito are like.

Given the fact that the approach the CEFR takes is language- and context-independent, we need to adapt the CEFR when applying it to specific local contexts. For example, we need to interpret and translate the proficiency descriptors into meaningful, contextualised, language- and learner-specific descriptors if we want to apply them to specific contexts, languages and purposes. The following graph

illustrates the necessary steps involved in making the CEFR context-specific for assessment purposes:

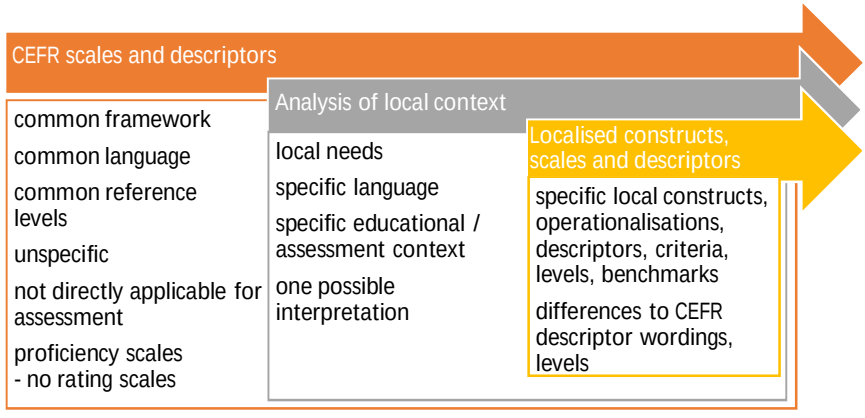


Figure 1. Making the CEFR context-specific.

As the largest (orange) box depicts, the CEFR as a language- and context-independent framework provides a common frame and a common metalanguage. It also provides illustrative descriptions of the common proficiency levels. The CEFR scales and descriptors serve the function of proficiency scales and cannot directly be applied for assessment purposes (cf. Alderson, 1991 for different scale functions). The descriptors do not describe assessment tasks or test characteristics, and they are not specific enough to serve directly as e.g. rating scales. When it comes to applying the CEFR to a local context (depicted in Figure 1 in the grey box in the middle), this context needs to be carefully analysed for local needs, specific language expectations, educational parameters including learner characteristics, as well as local assessment expectations and purposes.

Based on such an analysis of the local context, the CEFR can then be used as point of reference. The CEFR was not intended to provide a test blueprint that would specify what a test item targeting a given CEFR level should look like. Rather, relevant proficiency features for a specific assessment context have to be identified in the local context and "translated" into test specifications relevant for that context. Similarly, the CEFR cannot provide a description of language-specific features aligned to specific proficiency levels. Such descriptions will have to be developed in specific educational contexts for specific assessment purposes. Such a "translation" of the CEFR scales and descriptors into localised constructs, scales, criteria descriptors and benchmarks (as depicted in the smallest (yellow) box in Figure 1) will necessarily lead to differences between the local wording and the original CEFR wording. This is inevitable when the CEFR is to be adapted to fit local contexts, as was intended by the Council of Europe (Council of Europe, 2001: 7-8). Such differences in how the CEFR is interpreted and adapted, however, lead to the fact that CEFR levels and their operationalisations take on different interpretations and meanings in different contexts – resulting in the insight

that “your B1 is not necessarily my B1”. Suffice to state at this point that assessment providers can enhance transparency of local exams by documenting local adaptations, by communicating what the local CEFR interpretations look like and by demonstrating how the CEFR levels are locally operationalised – in other words, by sharing their “recipe for their interpretation of B1”. This is a prerequisite to enable a *common* reference to the CEFR levels and to communicate the localised meaning of the CEFR levels (see e.g. Harsch, 2014, 2018 for a more detailed argumentation).

3. One example for a localised adaption of the CEFR for educational monitoring purposes

I will now illustrate such a local adaptation of the CEFR for educational monitoring purposes by the development of national educational standards (NES, see footnote 3 above) in Germany. For Germany’s three school tracks, different standards were developed by a group of educators representing all of Germany’s regions and school forms. For the modern foreign languages, the NES are outcome-oriented, stating what learners at the end of secondary schooling are expected to be able to do with the language. The CEFR formed the starting and reference point, with its focus on learner language, its chapters on curricula and tasks, and with its proficiency scales informing the formulation of the standards. The proficiency model in the NES was modelled on the CEFR proficiency conceptualisation, and the core descriptors defining the NES were based (often verbatim) on relevant CEFR descriptors. In a second step, taking the NES as basis, tests were developed that aimed at monitoring the attainment of the NES at the end of the lower and middle school tracks. As with the NES, the CEFR served as a starting point to derive test specifications and constructs. Here, the CEFR chapters on tasks and assessment facilitated the test development project, and the CEFR’s proficiency scales were “translated” into test specifications. A group of teachers representing the German school system were trained to develop the specifications and the tests. Here, the CEFR helped enhance teachers’ understanding of test constructs and content, and their professional development, as is reported for other contexts as well (cf. Figueras, 2007).

The NES tests operationalise CEFR levels A1 to C1 to account for student proficiency above and below the NES. For writing, ratings scales were derived from existing descriptors based on the CEFR, and refined and validated in a combined training & revision process (Harsch & Martin, 2012). All tests were formally aligned to the CEFR (Harsch, Pant, & Köller, 2010). Hence the results of the large-scale assessment that is regularly conducted are reported on proficiency levels that are aligned to and derived from the CEFR (Köller, Knigge, & Tesch, 2010). Here, the CEFR-aligned localised proficiency levels serve to communicate attainments and educational monitoring to relevant stakeholders in the German school system. All interpretative steps are transparently documented in test specifications and publications; accompanying research has been published in books and journals; the

localised interpretations can thus be shared with a wider public audience and stakeholders in other educational contexts. Such documentation serves a similar function as the aforementioned Mojito recipe: Stating what ingredients, contents and procedures have gone into a Mojito adds transparency to what it might taste like, just as documenting test development derived from and aligned to the CEFR facilitates a shared understanding of the localised interpretation of the CEFR levels.

4. From localisation to a common reference point

The CEFR proficiency levels are intended to provide a common reference point, thereby facilitating the comparison of learning aims and teaching outcomes across different contexts. The CEFR is used as a common point of reference in many European educational systems, and it is referred to by most international exam providers. All major tests operating internationally have been aligned in one way or another to the CEFR. This alignment of curricula, educational systems and exams has on the one hand lead to greater comparability; on the other hand, it might evoke expectations of equivalence that in reality may be difficult to meet. For instance, different English proficiency tests aligned to the CEFR may be perceived by test users as measuring "the same", with the implication that the resulting classifications of test takers into CEFR proficiency levels should be comparable. This expectation is expressed, for example, by university admissions calling for test equivalence tables. However, this equivalence is not a necessary consequence from test alignment to the CEFR, for a number of reasons which the following passages set out to explain. The aim of the following section is to scrutinise why taking different tests aligned to the same framework may nevertheless lead to differing results.

4.1 The "rubber ruler"

To start with, measurement in the realm of language proficiency cannot be compared to measurement in the natural sciences. We do not measure hard facts such as temperature or length, which can be measured on a ruler, the units of which always stay the same. Douglas (2010: 3) fittingly introduces the picture of a "rubber ruler" to characterise what instrument we would need to measure language proficiency. As he explains (ibid., 3-4), the meaning of the units of measurement in language testing, be it the CEFR levels or units such as *beginner*, *intermediate*, *advanced*, is not precisely defined; the units are not equidistant and there is no absolute zero. Furthermore, a learner classified as being at level B2 (e.g. by a test score of 80 points in an imaginary test) is not "twice as proficient" as a learner at B1 (who scores 40 points in the same imaginary test). To make things more complex, re-taking a test will very likely result in a different score. Nevertheless, the "inexact" realm of language proficiency can be measured within certain limits of accuracy, because we can establish the measurement

error (which Douglas compares to the "stretch" of the rubber ruler, ibid.); we also can improve accuracy by using more measurement points. Moreover, there are means to compare different interpretations of the units of measurement or proficiency levels. It is in this last realm that the CEFR can be of great help, adding transparency to how a proficiency level can be described and interpreted. As outlined in Figure 1 above, the CEFR can inform test specifications, which in turn can inform stakeholders how the proficiency levels are interpreted in a local context and for a specific test or exam.

4.2 What does it mean to be at a level?

Exam providers should, therefore, carefully specify their tests with regard to what they measure, how they measure, and how they interpret the proficiency levels (e.g. the CEFR levels) they set out to measure and report. In the case of proficiency testing, a test usually results in statements about test takers' proficiency levels. Hence, one of the fundamental questions that also needs to be made transparent is what it means for a specific exam provider, a specific test and its test takers to be classified as having attained a certain proficiency level. The question of what it means to "be at a level", trivial as it may seem, is a complex one to which there is not one exact answer. Rather, we are again in the realm of the "rubber ruler", since the classification of test takers into proficiency levels involves human interpretation besides "hard" statistical analyses.

At this point, I need to briefly digress into the realm of statistics. Most internationally operating exams nowadays employ methods belonging to the so-called Item-Response Theory (IRT) to determine test takers' proficiency levels. IRT encompasses probabilistic ways of estimating test takers' overall proficiency levels based on their performance on the individual test items. IRT can simultaneously model test taker proficiency, item difficulty, as well as rater severity and assessment criteria difficulty for the productive skills. All facets are reported on the same IRT scale, thus directly showing the relationship between test takers, item difficulties, as well as rater harshness and assessment criteria. This IRT scale usually forms the basis for the endeavour to align a test to the CEFR.

Let us look at an example: An imaginary writing test with three tasks is to be aligned to the CEFR. After the tasks have been piloted and their quality has been assured, the tasks are administered to 200 learners, assessed by a local rating scale, and the resulting scores are subjected to IRT analyses. Figure 2 shows the outcomes in a simplified way: The three tasks are ordered according to their difficulty on the IT scale (left side in the figure), with task 1 being the easiest and task 3 being the hardest; the test takers[21] are distributed roughly in a bell curve along the IRT scale and the rating scale – the higher their ratings on the rating scale (right side in the figure), the further at the top of the IRT scale they are located, and the more proficient they are.

[21] Each symbol of a test taker in Figure 2 stands for 50 test takers in this example.

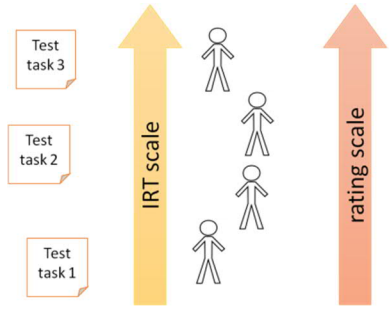


Figure 2. Item-Response Theory (IRT) scale.

One of the reasons why different tests of language proficiency may come to differing outcomes is that they operationalise slightly different constructs in different ways, and use different rating scales that may contain different assessment criteria. Such differences have recently been systematically analysed for the case of Flemish university entrance exams by Deygers (2018), who could attribute classification differences (learners being placed at different proficiency levels) to the different constructs, formats and assessment criteria used in the two tests he compared.

Going back to our imaginary test and its alignment to the CEFR, the next step after IRT scaling is the so-called standard setting phase (see the next section for more details), where a panel of judges evaluates the tasks and learner performances against the proficiency descriptions in the CEFR (depicted in Figure 3 below in the green arrow). The panel sets cut-scores, i.e., it decides about where on the IRT scale one CEFR level ends and the next one begins, as depicted by the (green) lines in Figure 3.

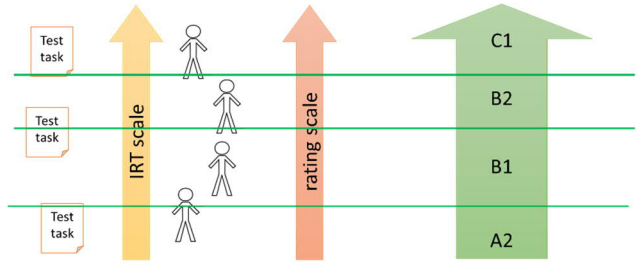


Figure 3. Aligning a test to the CEFR.

This process, however, contains a certain amount of uncertainty or inexactness, because the judges' decision most likely will not be unanimous – here lies another reason why different tests may lead to differing classifications of test takers. As is the case with any human judgement, different judges and panels may come to differing decisions. Once the boundaries (or cut-scores) are set, the next question to be answered is what test score in a given test is needed to be classified as "being at a level". This brings us back to IRT scaling and its implications for the probability of solving a task or an item

of a certain difficulty. Generally, the boundaries are set with the assumption that a person at the beginning of a level has a 50% probability of solving tasks and items that have a corresponding difficulty, i.e., that are located on the same point on the IRT scale. The probabilistic model behind the IRT scaling means that the easier the tasks or items get, the higher a person's probability is to solve the tasks; the more difficult a task gets, the lower the probability for that person to solve the task. This relationship is pictured in a very simplified way in Figure 4, which is inspired by De Jong (2004).

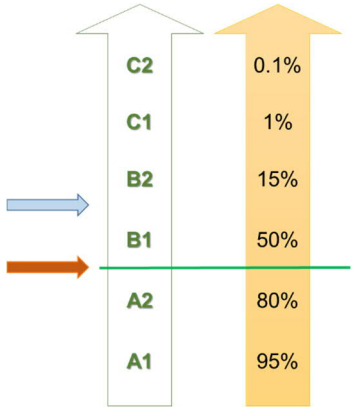


Figure 4. What does it mean to be at a level?

The (green) line depicts the beginning of CEFR level B1. A test taker located here (as a result of IRT scaling) has a 50% probability to solve items located at the beginning of B1, with the probability decreasing as the item difficulty goes up (e.g. to 15% response probability for items at level B2), and the probability increasing as the items get easier (e.g. an 80% probability to solve items located at A2). The question arises where a test taker should be located on the IRT scale so that we can assume with a sufficient amount of certainty that this person "is" at level B1. Should this person be located at the lower dark (orange) arrow, i.e., the beginning of B1, with the implication that this person can solve A2 items with an 80% probability and B1 items with 50% probability and falling? Or should the person rather be located towards the top end of B1, indicated by the upper light (blue) arrow in Figure 4, so that we can assume that the test taker has an 80% probability of solving the tasks and items located at B1?

This question is indeed answered differently by different researchers and exam providers: Some exam providers require that test takers have a 50% probability to solve the items of a level to be classified as being at that level, some set this probability higher. If it is requested that a person has to be able to solve the majority of the items of a level to be classified as being at that level, the response probability for items located at the beginning of the level must be much higher than 50%. These different interpretations of what it means to be at a level are yet another reason for discrepancies between different tests. In our imaginary case, if the exam provider operates with the 50% probability, a test taker located at the orange arrow would be classified as being

at B1. However, if the exam provider requests that the test taker should be able to solve the majority of the items of a level, the same test taker would have been classified as being at A2. Thus, different interpretations of what it means to be at a level are another reason for the non-equivalence of different tests that are aligned to the CEFR.

4.3 Different alignment procedures – differing outcomes

So far, we have established three reasons for non-equivalence of tests being aligned to the same framework, in our case the CEFR: the exam providers' differing interpretations of the CEFR with regard to their constructs, test tasks and assessment criteria; the exam providers' differing interpretations of what it means to be at a level; and discrepancies among human judgements in the standard setting phase. Let us now examine the latter phase more closely, as its inherent reasons for non-equivalence are quite complex.

Alignment to the CEFR is meant to increase test transparency and ultimately can add to comparability. There is a body of literature that test developers and exam providers can refer to, such as the Manual (Council of Europe, 2009) or reports like the ones by Figueras & Noijons (2009) or Harsch et al. (2010) for research on standard setting endeavours in Europe. The Manual outlines possible steps and procedures of specifying and aligning test content and outcomes to the CEFR. It encourages increased transparency on the part of test developers, provides practical tools and is complemented by technical supplements. It describes a range of commonly used standard setting methods that are deemed suitable for aligning tests to the CEFR. Yet following the recommended procedures and applying formal standard setting methods alone does not automatically lead to test comparability, let alone equivalence of outcomes of different tests, for the reasons outlined above, and also because these methods come with their own uncertainties. What alignment and standard setting do help with is making transparent the relation between a (localised) test and the (common, generic) CEFR levels.

I will now illustrate some of these uncertainties with reference to three well-documented and often used standard setting methods, i.e. Angoff, Bookmark and Basket. The Angoff Method requests the judges to estimate the probability for each test item that a 'borderline candidate' (at the boundary between two adjacent proficiency levels) can answer the test item correctly. The Bookmark Method presents all items in ascending order of difficulty and asks judges to virtually 'place a bookmark' between the last item a borderline candidate would be able to handle and the first item deemed too difficult for such a candidate to solve. In the Basket Method, judges have to determine at what proficiency level a candidate minimally has to be to be able to answer a test item correctly. All three methods pose the same three main challenges to the judges. First, judges have to imagine a hypothetical borderline candidate; research indicates that different judges might refer to different interpretations of such a candidate (e.g. Harsch & Hartig, 2015). Second, humans are not very apt at judging probabilities (e.g. Kahneman, 2011). Third, as indicated above, judges may not have a

shared understanding of what 'being at a level' means, and fourth, there are inherent error margins in human judgement. Furthermore, it is known that different standard setting methods yield differing results, and the composition of the panel influences the outcomes, i.e. the setting of level boundaries and pass scores depends on the judges and the methods.

In sum, test alignment and formal standard setting contain certain uncertainties, similarly to the uncertainties I have outlined above for measuring language proficiency. Hence, it is of utmost importance to transparently document the constructs, contents, operationalisations and approaches taken, as well as alignment procedures, standard setting methods and the test provider's understanding of what it means to be classified to be at a level. This serves to report the localised (i.e., the test developer's) interpretation of the CEFR levels and the test provider's rationale behind their placing test takers at CEFR levels. Documenting and publishing these decisions serves to add yet another layer of transparency to the aforementioned "Mojito recipe".

5. Reporting on a common scale does not imply test equivalence

When it comes to reporting test results of a test that is aligned to the CEFR, the 'local' test scores are usually also reported with reference to the CEFR proficiency levels. Exam providers usually publish their alignment endeavours along with research on the reliability and validity of the standard setting procedures, and they usually publish score alignment tables, stating the range of (test or band) scores that align to a certain CEFR level, as for example is done for the Cambridge Main Suite (UCLES, 2015) in relation to the Cambridge English Scale and the CEFR:

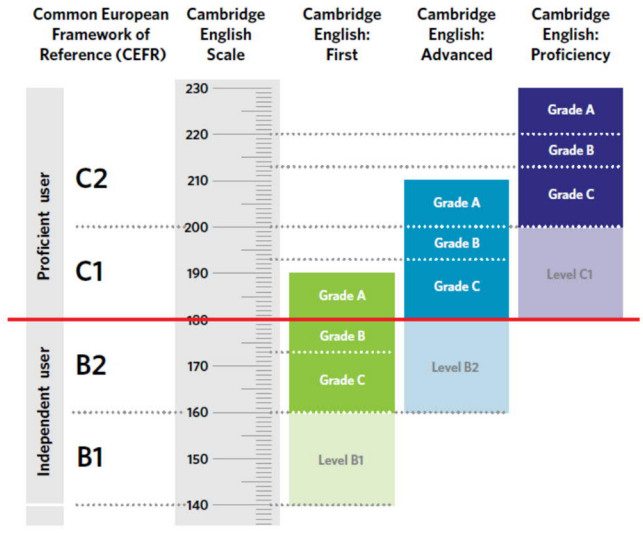


Figure 5. Cambridge Main Suite and Cambridge English Scale aligned to CEFR (UCLES 2015: 4, online: www.cambridgeenglish.org/in/exams-and-tests/cambridge-english-scale; the red line indicates the cut score between B2 and C1).

Another example for such an alignment table is found for the Pearson Global Scale of English (GSE). The GSE was developed for reporting the PTE Academic, by mapping the test scores on the IRT values of the original CEFR descriptors[22]; the outcome was validated by formal standard setting procedures (De Jong & Benigno, 2017).

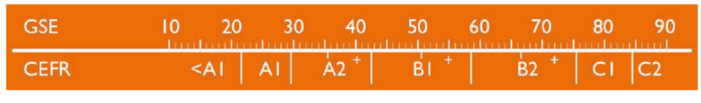


Figure 6. Alignment of the Pearson Global Scale of English to the CEFR (ibid.: 5).

Such alignment tables help test users to quickly establish the link between (local) test scores and CEFR levels. By referring to the CEFR proficiency scales and their underlying descriptors, test users can get qualitative feedback about the meaning of the test scores, i.e., what a test taker with a certain test score is likely to be able to do with the language. Yet one has to bear in mind that these tables do not contain any information about test purposes, content, construct or scoring procedures, and they do not help test users in deciding whether a test is appropriate for their local context.

Often, test users want to compare different tests, For this purpose, some exam providers publish direct comparison tables between their own and other tests, as for example ETS does with an online tool for a comparison of TOEFL iBT® and IELTS, based on a comparison study (ETS, 2015):

[22] The CEFR scales were developed using IRT scaling (North, 2000).

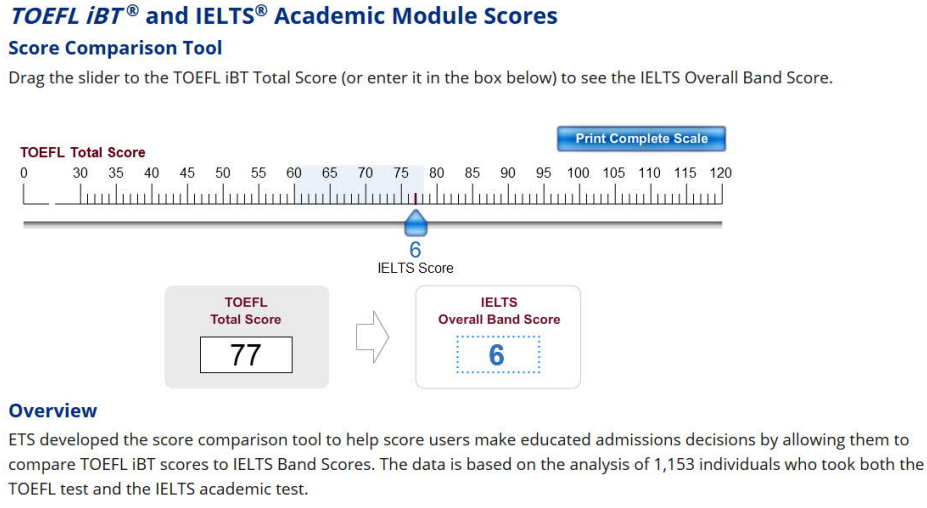


Figure 7. Score comparison tool, source: https://www.ets.org/toefl/institutions/scores/compare, accessed 18.05.2018.

This figure evokes a precise ruler that is not "rubbery", in seeming contrast to what I have argued above. This may be justified in as far as the data analysis is based on "hard facts", because, as the webpage states, the comparison tool "is based on an analysis of 1153 persons who took both" tests. Yet, while this tool may help test users with quickly converting test results, we do not know how comparable the constructs, the tasks, or the scoring criteria of the two tests are. Moreover, the two tests use very different means of reporting, with ETS reporting scores from 0 to 120, and IELTS reporting much coarser band scores (from 0 to 9, with 0.5 steps in between), and the tool does not state the tests' measurement errors. Hence, while we get an idea of how the reported results from the two tests correspond, we cannot assume that the two tests measure the same (see also Deygers, 2018, for the Flemish context), nor do we know any details on the exam providers' interpretation of the underlying proficiency levels.

The question arises whether the CEFR could serve as a common reference point and means of comparison, because all major internationally operating English proficiency tests are aligned to the CEFR. Yet this is all but a straightforward endeavour. First, for the reasons outlined above, there is a certain amount of uncertainty and measurement error in any alignment to the CEFR. In addition, the CEFR levels are rather broad. If we now take the measurement error into account that is inherent in any test, this adds to the inexact nature of aligning scores to the CEFR. This inexactness is exacerbated if test results are reported as score bands, as these bands add to coarseness and hence to inexactness.

The effect of measurement errors and the resulting uncertainty is illustrated nicely by a comparison study published by De Jong & Benigno (2017). They compared test dimensionalities and score reliabilities across PTE Academic, TOEFL iBT and IELTS, by means of referencing the scores from these three tests to the aforementioned

Global Scale of English (GSE). In their reliability study, they compared the errors of measurement of the three tests for a certain score range on the GSE (the GSE functioning thereby as a proxy to the CEFR) and found a substantial score range for all three tests[23]. This means e.g. for PTE Academic that the true score of a test taker with a score of 59 lies in the range of 54-64 scores (ibid.: 12). If one now looks at the alignment of Pearson's proficiency scale to the CEFR (see Figure 6 above), that range covers pretty much the whole range of CEFR level B1. So we can say that a person with a score of 59 is most likely to be around the middle of B1, but we cannot be absolutely certain. This level of uncertainty is not unusual and is found with all tests – De Jong & Benigno (ibid.) actually found higher score ranges for TOEFL and IELTS than for PTE Academic. We could do this exercise for any proficiency test aligned to the CEFR and would get similar results regarding the precision (or degree of certainty) of aligning scores to CEFR levels. What is very welcomed for test users is that exam providers transparently publish data on measurement error in a way that is also accessible for laypersons.

To sum up, because different tests use differing scoring and reporting systems (also in terms of their coarseness), have differing errors of measurement, and use different avenues to align their scores to the CEFR levels, it is difficult to produce a "hard and accurate" alignment table where the different test scores are mapped to the CEFR in such a way that a direct comparison between the tests becomes possible. While De Jong & Benigno (2017, p.17) or ETS (2015), for example, present such research-based comparison tables, these tables represent the view of one exam provider, which is not necessarily shared by the other exam providers. Test users such as university admissions are well advised to use all available sources when compiling their own tables, triangulating existing alignment and comparison studies, and bearing in mind that even in cases where direct comparison data exist, these data do not necessarily reveal whether the different tests yield acceptably similar classifications of test takers to CEFR levels (see e.g. the results from Deygers, 2017; 2018).

6. Not all is lost – how test results can meaningfully be compared

This rather bleak picture is by no means a reason for despair. It is rather a reason to close the circle and go back to the local context in which the tests are to be used: Test users need to take into consideration the purpose they want to use the test for, their target group and local language requirements. As a starting point, local needs analyses are recommended, particularly with regard to establishing the appropriate language requirements in terms of CEFR levels (see e.g. Abdulhaleem & Harsch, in print) for an example in the Saudi Arabian higher education context). Next, there are test-specific

[23] The score range refers to the range within which the "true score" of a test taker is expected with a 95% level of certainty. The true score of a test taker is the hypothetical score a test taker would get if there was no measurement error; the actual reported score is called the observed score. The true scores lies within 2 units of the standard error of measurement around the observed score.

questions that should be addressed by test users, such as: Does the test's construct, content, criteria, formats, and scoring approaches fit the local needs? How trustworthy are the scores and the alignment to the CEFR, i.e., what measurement errors are reported and what do we know about the alignment procedures and outcomes? Then there is the realm of predictive validity studies to be addressed, i.e., studies that examine whether the test scores are meaningful predictors of how well the test takers are prepared for the language requirements of the local context.

Needless to say that such needs analyses and predictive validity studies require resources and expertise that not all test users have at their disposal. Here, many of the major test providers offer research grants to pursue such studies in local settings, adding to the body of knowledge about the appropriateness of certain tests and their score reporting systems for specific local contexts and settings.

7. Why detailed score reports matter

One aspect that deserves closer attention is the level of coarseness in score reports, as it has the most direct implications for test users. Some tests report a fine-grained score profile for different skills on a numeric scale, others provide one overall score band that encompasses a certain range of raw scores. From the perspective of test users, the more details a report contains, the more information it provides. This has implications for decisions that test users make with reference to the CEFR levels. If all I receive is a coarse band score that is aligned to a broad CEFR level (or, for that matter, a fine-grained test score with a large measurement error), then I only know roughly where a test taker may be, e.g. somewhere around the middle of B1. If, however, a test report shows where on the CEFR level a score is located and if the test has a reasonable measurement error, I get a more precise picture whether a test taker is most likely at the beginning, middle or upper end of a proficiency level.

Coarseness in reporting is also the last reason for differing CEFR classifications by different tests that I want to address here. If a test taker is classified by one test as being at B1, and by another test as being at B2, the two tests may actually report a very similar result, i.e., the test taker may actually be a borderline candidate. Here, we have to bear in mind the fact that the boundaries between the proficiency levels are set with a certain degree of uncertainty, as explained above. It is worth noting that a person classified as being at the upper end of B1 may in fact be much closer to a person at the beginning of B2 than to a person who is also classified as being at B1 but is actually at the beginning of the level, as Figure 8 depicts.

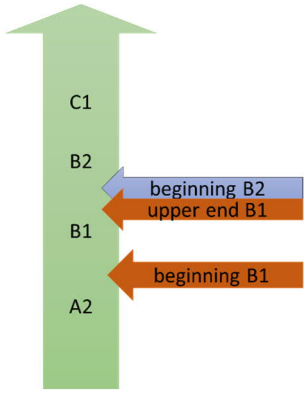


Figure 8. Implications of score report coarseness for CEFR level classifications.

The coarser the score reports are (or the larger the measurement error), the less precise the information will be; more fine-grained score reports (and smaller measurement errors) allow a clearer picture of where on the proficiency level the test taker most likely is located. In addition, if the reports are broken down for different skills, they provide fine-grained feedback that is more informative for test users than one overall grade.

8. Conclusions and ways forward

In sum, the CEFR provides a reference framework and a meta-language to communicate different aspects of language learning, teaching and assessment, and to increase the transparency and comparability of curricula, educational systems and exams – but it was never intended to provide a common "measurement" scale that would allow a direct comparison of different test scores. Neither was it designed as a tool that would ensure test equivalence of different tests. In this chapter, I attempted to unmask the assumption of test equivalence as unreasonable for a variety of reasons that lie in the inherently imprecise nature of assessing language proficiency, in the imprecision of human judgements, and in the necessarily differing localised interpretations of the CEFR levels and their operationalisations. I argued for accepting that ultimately, while our interpretations of the CEFR levels will differ, we can share our "recipes" of what our local interpretations and operationalisations are like. In order to enhance comparability and transparency, I can only reiterate the need for detailed documentations of such interpretations, along with test specifications and standard setting reports. Next, I would like to stress the importance of a transparent test reporting and feedback system, so that test results are reported with as much details and precision as possible with regard to where test takers are and what areas they need to improve. The reports should ideally state the alignment of fine-grained scores to CEFR levels, along with the measurement error.

In order to improve test comparability and help test users select the most appropriate tests, local needs analyses are required, more alignment and comparison studies between different tests are needed, as well as more predictive validity studies in local settings. Here, test users and exam providers are asked to collaborate, as well as researchers, in ensuring that tests are used in local contexts in appropriate ways. Related to this call for closer collaboration is the need to foster assessment literacy amongst all stakeholders, i.e. develop the necessary knowledge base, skills and competences to make informed and justifiable decisions. Here, professional associations such as the European Association for Language Testing and Assessment EALTA (ww.ealta.eu.org) or the International Language Testing Association ILTA (www.iltaonline.com) provide workshops, webinars and conferences, along with online resources. The Association of Language Testers in Europe ALTE (www.alte.org) represents test providers and offers courses and conferences, as well as a quality auditing system of European language examinations. Some test providers offer workshops for stakeholders wanting to use their tests. There is an expanding body of research in the realm of assessment literacy, along with a growing body of introductory literature (e.g. Douglas, 2010; Fulcher, 2010; Green, 2014). All these endeavours support the development of expertise in using language tests and score reports in a fair, reliable and valid way.

References

Abdulhaleem, E. & Harsch, C. (2018). Using the CEFR Scales to Assess Students' Proficiency Levels in a Saudi-Arabian Higher Education Context. In Brandt, A., Buschmann-Göbels, A. & Harsch, C. (eds). *Der Gemeinsame Europäische Referenzrahmen für Sprachen und seine Adaption im Hochschulkontext.* Fremdsprachen in Lehre und Forschung Bd. 51. Bochum: AKS Verlag, 167-178.

Alderson, J. C. (1991). Bands & Scores. In J. C. Alderson & B. North (Eds.), *Language Testing in the 1990s* (pp. 71–86). London: Macmillan.

Alderson, J. C. (2007). The CEFR and the Need for More Research. *The Modern Language Journal*, *91*(4), 659-663.

Van Avermaet, P. (2004). *Is my Mojito your Mojito?* Paper delivered at ALTE Conference Day, ALTE meeting, Bilbão, Nov. 2004.

Broek, S. & van den Ende, I (). *The implementation of the Common European Framework For Languages in European education systems.* Brussels: European Parliament, Policy Department B: Structural and Cohesion Policies, available online: http://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/495871/IPOL-CULT_ET(2013)495871_EN.pdf, accessed 26.10.2018.

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

Council of Europe (2009*). Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A manual*. Strasbourg: Language Policy Division, available online: https://www.coe. int/t/dg4/ linguistic/Manuel1_EN.asp, accessed 18.05.2018.

Council of Europe (2018). Common European Framework of Reference for Languages:

Learning, teaching, assessment. Companion volume with new descriptors. Strasbourg: Language Policy Division, available online: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989, accessed 18.05.2018.

Deygers, B. (2018). *University entrance language tests: examining assumed equivalence*. In J. Davis, J. Norris, M. Malone, T. McKay, & Y Son (Eds.). Useful Assessment and Evaluation in Language Education. Washington, D.C.: Georgetown University Press, 2018.

Deygers, B. (2017) Assessing high-stakes assumptions. A longitudinal mixed-methods study of university entrance language tests, and of the policy that relies on them. Leuven, Belgium: KU Faculteit Lettere.

De Jong, J. (2004). *What is the role of the Common European Framework of Reference for Languages: Learning, teaching, assessment?* Paper delivered at EALTA Conference, Kranjska Gora, May 2004.

De Jong, J. & Benigno, V. (2017). Alignment of the Global Scale of English to other scales: the concordance between PTE Academic, IELTS, and TOEFL. Pearson: Global Scale of English Research Series. Available online: https://prodengcom.s3.amazonaws.com/GSE-Alignment-other-scales.pdf, accessed 18.05.2018

Douglas, D. (2010). *Understanding Language Testing*. London: Hodder Education.

ETS (2015). *Linking TOEFL iBT ™ Scores to IELTS® Scores – A Research Report.* Available online: http://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf, accessed 18.05.2018.

Figueras, N. (2007). The CEFR, a Lever for the Improvement of Language Professionals in Europe. *The Modern Language Journal, 91*(4), 673-675.

Figueras, N. & Noijons, J. (Eds.). (2009). *Linking to the CEFR levels: Research perspectives.* Arnhem: CITO, CoE, EALTA.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.

Green, T. (2014). *Exploring Language Testing and Assessment*. London: Routledge.

Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly, 11*(2).

Harsch, C. (2018). How suitable is the CEFR for setting university entrance standards? *Language Assessment Quarterly, 15*(1).

Harsch, C. & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly, 12*(4).

Harsch, C. & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17*(4).

Harsch, C., Pant, H. A. & Köller, O. (2010). Calibrating standards-based assessment tasks for English as a first foreign language. Standard-setting procedures in Germany. Münster: Waxmann.

Hilden, R. & Takala, S. (2007). Relating descriptors of the Finnish school scale to the CEF overall scales for communicative activities. In A. Koskensalo , J. Smeds , P. Kaikkonen and V. Kohonen (eds), *Foreign languages and multicultural perspectives in the European context.* Vol. 9-10, Lit Verlag, Berlin, pp. 291-300. Available online: http://hdl.handle.net/10138/26407, accessed 26.10.2018.

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus and Giroux.

Köller, O., Knigge, M. & Tesch, B. (2010). *Sprachliche Kompetenzen im Ländervergleich. Überprüfung der Erreichung der Bildungsstandards für den Mittleren Schulabschluss für Deutsch und die erste Fremdsprache in der neunten Jahrgangsstufe*. Münster: Waxmann.

Little, D. (2005). The Common European Framework and the European Language Portfolio: involving learners and their judgements in the assessment process. *Language Testing, 22*(3), 321-336.

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency.* New York u. a.: Lang.

Rupp, A.A., Vock, M., Harsch, C. & Köller, O. (2008). *Developing Standards-based*

Assessment Tasks or English as a First Foreign Language – Context, Processes and Outcomes in Germany. Münster: Waxmann.

UCLES (2015). *The Cambridge English Scale explained*. Cambridge: UCLES, available online: www.cambridgeenglish.org/images/177867-the-methodology-behind-the-cambridge-english-scale.pdf, accessed 18.05.2018.

Weir, C. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281-300.

Digitally testing the language of young learners: A learning curve

Angela Hasselgreen and Eli Moe
University of Bergen

1. Introduction

It was 2002 when the phone call came through from our Ministry of Education. They wanted us to be responsible for making digital tests of English for pupils at key stages in the Norwegian school system, preferably linked to the CEFR (Common European Framework of Reference for Languages, Council of Europe, 2001). The primary purpose was to report back to authorities, both national and local, on the level attained by pupils, although, more recently, the formative purpose of the tests, through feedback to teachers, has been given a growing role. The tests are not intended to be used in the setting of formal grades.

'We', at that stage, were the two authors of this article – one with experience in low-stakes testing and assessment of the English language of school children, and the other with experience in high-stakes, nationwide testing of the Norwegian second-language ability of adults. We were both ex-teachers of English and had both worked with the CEFR/ELP. While we could handle word processing, what we knew about the use of computers in testing was limited to experience in the European Union's *DIALANG* Project (1996-2004).

Notwithstanding the huge challenges ahead, we were eager to take on the task. Under the auspices of the University of Bergen, a group had to be put in place consisting of some colleagues with similar backgrounds to our own, including one with experience in computer-assisted language learning. We quickly added into the mix a current primary school English teacher and that essential item for young learner testing, an illustrator. A digital media group allied to the University agreed to shoulder the technical challenges. Yet, high in enthusiasm, and well-off for resources, we still lacked confidence in the face of a task of a nature and magnitude that no institution we knew of had faced at that time. Our response was then – as in a smaller way in the past – to call on Sauli Takala. Typical of Sauli, and busy as he always was, he unhesitatingly and very generously agreed to advise and guide us as we embarked on our daunting journey. One of his first moves was to bring Felianka Kaftandjieva on board our team. She was indispensable as both statistician and friend for the rest of her life.

Sauli kept an avuncular eye on us throughout the early years of our project, offering quiet wisdom and seeming to enjoy the rather mad humour that has always

pervaded our work. He weaned us gradually, and a decade and a half on, we have grown in confidence and stature, but are still on that eternal learning curve. In this article we will consider some aspects which have been central to the development of our testing projects: the learners themselves, the tests we make, and finally ourselves – the testing group and the basic working method we have established.

2. Young language learners

The YLs (young learners) specifically to be addressed in the National Test project were pupils at four key stages: 9-10, 12-13, 15-16 and 16-17 years. Two non-compulsory tests, for pupils around 8-9 and 16-17 years, were subsequently added to our repertoire. These are intended to be used by teachers in order to provide insight into the level of English of individual pupils, on a range of skills. Thus, it has been an essential part of our work to be aware of the characteristics of YLs across a range of stages, and of what can and should be asked of them in language testing.

Here we will briefly consider aspects of YLs which from the outset were seen as most relevant to our work, specifically cognitive and social development (affecting task types), actual language elements/themes testable (affecting test content) and computer skills (affecting format). We will also look at some findings on L1 (first language) development which influenced us underway and will finally bring these elements together in a brief consideration of relevant CEFR levels.

Since the early works of Piaget (e.g. 1926), almost a century ago, stages in the cognitive and social development of children have been the focus of many studies. Based on our reading (e.g. Cameron, 2001), backed up by intuition and experience as teachers, we were aware of many characteristics of YLs that were salient to our work.We were aware that the youngest pupils have a limited attention span and a great need for play, fun, games and fantasy. We knew that their world knowledge is largely based on concrete personal experience, and that they are relatively egocentric.

We also knew that older children have longer attention span and are better able to cope with abstract ideas and problems requiring simple logic. We were aware that they are learning to collaborate and be more aware of others. We learnt that they can carry out more complex tasks and create a ‘wholeness’ from parts, getting the gist of information. We were aware that this development continues through the teenage years, with an ability to cope with increasingly abstract, complex and ‘remote’ ideas.

These characteristics offered us an insight into the kind of tasks we were able to design for pupils, thus taking into account Cameron’s (2001:25) warning that the demand of tasks go beyond the linguistic. Our in-house teacher was invaluable in voicing concern for the younger children, regarding the overall design of tests, e.g. their dependence on pictures, their inability to sit still for long periods and their anxiety if confronted with items they could not manage.

While we had a good idea of what children were familiar with through their English learning, and had general guidance from the school curriculum, we needed to be sure that our more basic test items would not present content that some children had

not met. We needed to know which specific themes we could assume they were familiar with, and what actual vocabulary this entailed. This involved consulting teachers, through a survey, as well as combing through the most commonly used course books, so that the final result was a glossary of themes and vocabulary.

When it came to the computer skills of pupils, we knew very little. This was of immediate concern, as the tests were to be digital, and in the early millennium years, even many adults were struggling with the mysteries of computing. Preliminary trials with a number of task formats in local schools allayed our worries to a large extent. However, it took a visit from the Ministry of Education to our labs, with local children invited in, to persuade our bosses that the children actually took to the tasks with great confidence, often outperforming the adults!

Thus, it was that we felt able to embark on the National test project, in 2003. In the years that followed we were continually learning, through local trialling and national piloting, what children could and, seemingly, could not do. We were also influenced by research findings, not least those of Nippold (2007) on the development of the first language in children and teenagers.

Nippold's meta-analysis of a wide-ranging body of L1 research findings is briefly reported in Hasselgreen and Caudwell (2016:6-12). Some of the conclusions we have found potentially relevant to our testing projects concern the lexicon, syntax and discourse, reasoning that, while it is difficult to predict what a child at any stage will manage in an L2 (second or foreign language), a ceiling may be set by what they can manage in their L1.

The lexicon: It is only from about 11 years of age that children appear to acquire a range of abstract nouns, or to understand figurative meanings alongside physical meanings of words such as *bright*. Between about 9 and 14 years, the understanding of derivational morphology, with affixes, such as *un-* and *–ness*, develops; this is believed to correlate highly with reading ability.

Syntax: The ability to link phrases and clauses within a sentence develops gradually throughout the school years, with an increasing range of conjunctions. Some, such as *although* and even *but* are believed not to be fully mastered by the age of 12. Links between sentences, using adverbial conjuncts such as *However*, is only mastered to a very limited degree before adolescence.

Discourse: While children around the age of 10 appear well-able to produce narratives, it is not before adolescence that a clear ability to use genres which take another person's perspective into account, such as persuasion or negotiation, is achieved.

On the basis of the aspects considered above – cognitive/social development, language domain, and age-related L1 ability – it is possible to posit some correlation between age and the approximate level on the CEFR that a YL might maximally be able to reach. Hasselgreen and Caudwell (2016) carried out such an analysis, and their results are presented in Table 1.

Table 1. Correspondence between age groups and CEFR levels potentially attainable.

Age groups	Limits of CEFR levels potentially attainable
Young children (roughly between 5/6 years and 8/9 years)	A2 Reading and writing levels will depend on the emergence of literacy.
Older children (roughly between 8/9 years and 12/13 years)	B1
Teenagers (roughly between 13 and 17 years)	B2
Exceptional older teenagers	C1

From Hasselgreen and Caudwell (2016:34)

It must be emphasised that the levels in the table are to be regarded as 'ceilings' for L2 ability, rather than what a child at a certain age can be expected to reach, as this is highly dependent on factors such as the learning environment. Subsequent standard setting on our reading test items indicate that, in the case of school children in Norway, the average level is around lower A2 for 5th grade. It is also estimated that the average level of achievement is around lower B1 for 8th grade.

3. The tests

The test development project we were originally assigned involved the National Test of English at the end of 4th. grade, 7th. grade, and, for a short initial period, 10th and 11th grades. Two tests were to be developed: reading (digital) and writing (non-digital). Similar National Tests were developed for numeracy and Norwegian literacy.

The writing test for each grade consisted of three independent tasks, with an increasing level of complexity. Teachers rated the tasks on a scale adapted from the CEFR, originally developed for children in the AYLLIT (Assessment of Young Learner Literacy) Project (Hasselgreen et al, 2012). Key teachers from districts throughout Norway were trained centrally, and they, in turn, trained all the relevant English teachers in their districts. This was demanding of resources, and inevitably led to rather low test reliability; these were two main factors in the decision to drop these tests after two years, when a reviewing break was put into effect for all the National testing. This was regrettable in many ways, as there was unquestionably a 'lift' in the assessment ability of teachers, particularly at primary school, where many teachers had no specialism in teaching English. Teachers were of course free to use the scale to guide them in their own classroom assessment, and anecdotal evidence suggests that this happened. The tests and rating scales had been quite closely monitored and influenced by Sauli Takala, and therefore examples of both are shown in Appendix 1.

The reading test, after the reviewing period, was re-established in 2007 for pupils at the start of 5th. and 8th. grade, i.e. 9-10 and 12-13 years. Our mandate was that the test should be digital and carried out online during a two week 'window' for each grade. Marking would be automatic, with results relayed to schools, with each pupil placed on a descriptor band: 1-3 for 5th. grade and 1-5 for 8th. grade. The tests should give all pupils the chance to demonstrate what they could do, rooted in the aims of the school curriculum. As a more formative purpose for the test gradually gained in importance, it was decided that each item be coded according to the particular aspect it measured. This decision had a great impact on our task as item-writers and posed what was probably our most significant challenge since that of deciding item formats. These two challenges, and how we have striven to meet them, will be the focus of the rest of this section.

3.1 The test format

At the very outset of our project, we were given an excellent piece of advice by a colleague with more computing experience than most at that time: do NOT start by designing paper-and-pencil items and transfer these to a computer, but rather think computer from the start.

Thus, it was that we shook off what we had traditionally thought of as test item formats and threw ourselves into the world of click and drag. The age of the pupils, particularly the 5th. graders, made pictures an essential ingredient of a high proportion of items; beside the visual effect of these, they were well-suited to depicting the more concrete concepts best coped with by this group. Pictures and texts therefore comprised the main elements, with occasional graphics such as tables. Consideration had to be given to limits imposed by the size of a computer screen and the fact that there was to be a one-to-one relationship between item and screen: every item had to fit into one screen, and any text that was used in more than one item, e.g. with a series of questions, had to be displayed with each question in a separate screen. Generally speaking, these issues were the terrain of the technicians. Our main task was to decide what operations pupils were to carry out.

As most tasks in our tests had to be 'closed', since the computer could not 'mark' free stretches of writing, a concern was to maximise the number of choices that pupils would have. Some items, based solely on texts, inevitably involved multiple-choice questions, where the pupils click on the correct alternative. A principle we held to was that four choices should be offered – fewer exposed the item to guessing, while more options increased the risk of implausible alternatives. This was also the case when gaps in a text had to be filled with a word (often a grammatical form) from a drop-down box. Similarly, items 'matching' texts with pictures either involved a text and four pictures, or four texts and one picture; this was a restriction imposed by the physical size of an item relative to the screen. However, we gradually became adept at creating formats that offered a wider range of options. These included clicking an object

or person in a picture and selecting and dragging an item into a specific location in a picture, making the tasks unlikely to be answerable by sheer guesswork. When text only was involved, more open items were created by clicking on a vocabulary item in a text ('Click on the word that means almost the same as..*) or by clicking on a name, to identify a person ('Who could say …'). The final 'closed' format we devised was 'Move paragraph' (8th. grade only), whereby a text of about five paragraphs was presented with the first in place, and the remainder jumbled, with vertical arrows to move the paragraphs up and down. The very small number of open items offered, testing grammar, involve pupils having to write a number of words in a gap. These have to be perfectly accurate, including spelling, to be counted as correct.

These formats were not all in place from the start, but emerged gradually, with other ideas being tried out and discarded underway. Thus, our tests currently have the following formats:

- Click picture (to 'match' a text)
- Click text (to 'match' a picture)
- Click and drag (moving an item into a picture)
- Multiple choice questions on a short text (approx. 100 word) or long text (approx. 300 word)
- Fill gap in text with word/word form (from drop down box)
- Click on the name of a person (who could say ..)
- Click on a word in a text (select a vocabulary item synonymous with a given word) (8th. grade only)
- Move paragraph
- Write words in gap with accurate spelling and grammar (8th. grade only)

Appendix 2 shows examples of some of these formats.

Aspects to be measured by items/coding

From the outset of the restoration of the tests, the developers were required to 'label' every item according to what it measured. And as the role of formative assessment became dominant school policy, it became increasingly expected that the tests should play a part in this, through feedback on pupils' test performance. Therefore, it was decided that a grid should be made available to teachers, showing what every item in a test measured, and allowing teachers to see how pupils had performed on individual 'subskills', in order to work with pupils to build these skills. This responsibility alarmed us to a certain extent, as it is well documented (e.g. Alderson, 2001) that there are many ways of classifying skills, sometimes regarded as 'strategies', and there is little evidence in an answer to indicate exactly which skills a reader is 'using' in order to arrive at the answer. Moreover, we felt that, statistically, there was a need for caution in drawing conclusions on the basis of a handful of items in a test which might target a particular skill.

Notwithstanding these reservations, however, we had to find a way of meeting these demands. Initially, we labelled items after a test set had been put together, in a fairly intuitive way. We knew that some items required finding *detailed information*, while others tested finding the *main point* in a text. We had consciously made items to test a word or phrases in a very short text match a picture and labelled these *vocabulary* items. This rather ad hoc set of labels was largely assigned to items on the basis of what we intended they should measure, rather than on any scrutiny of what they might actually measure. This was apparent in the case of items such as those consisting of four texts, each of several sentences, to be matched with a picture. While it was expected that these items would require a lot of detailed reading, some proved to be very easy, as the clue lay in a single word or phrase.

The development of a more stringent coding system, based on a consideration of what was felt to be actually necessary in order to answer an item, arose from a series of small investigations into features of our tasks as predictors of difficulty. The process required trained raters (our co-workers) to assign codes to items, which had known p-values, indicating the percentage of pupils who had answered the item correctly in the test. The initial attempts at this involved a rather unwieldy one-dimensional set of codes representing very different types of features, some relating to texts, others to tasks and some to the interplay between both of these. Items could be assigned several codes simultaneously, the analysis of such a mixed bag of items made it difficult to compare related features, leading us to the conclusion that we needed separate dimensions, whereby a single code on each dimension would be allocated to each item. Besides a mechanical measure of text length and readability index, two distinct dimensions were identified. The first of these dimensions involved an interplay between task and text, concerning the level of reading processing required to answer the task; this could involve understanding a single word or a sentence, or making links across the text.

The other dimension involved features most essential to the task itself – what the pupil had to 'do', such as to find detailed information or main point, or perform a grammatical operation. In all, 93 items with p-values based on over 50,000 test takers were coded by seven trained raters. It is beyond the scope of this article to present the statistical process, which is covered fully in Hasselgreen, Grocott and Torsheim (2017), but some tendencies will be reported. What is of main interest here is the coding system that emerged, which became the foundation for the reporting grid for teachers, as well as forcing us as item writers to consider what an item was likely to be measuring.

Dimension 1 is founded on Khalifa and Weir's (2009) levels of reading processing, and can be regarded as a hierarchy, with each level building on the one preceding it, as follows (from Hasselgreen, Grocott & Torsheim, 2017:223):

- *Vocabulary*: understand vocabulary – understand a word or phrase, possibly with the support of the context
- *Sentence:* understand sentence(s) – understand a sentence/clause, or a number of adjacent sentences/clauses

- *Link*: link sentences/parts of the text – make the connection between sentences which are separated in the text. This can also involve linking between different types of text, e.g. diagrams.

In coding items in this dimension, the rater had to decide on the highest level of processing required to answer the item correctly.

Dimension 2 involves the operation required by the task, and consists of five mutually exclusive categories (from Hasselgreen, Grocott & Torsheim, 2017:223):

1. *Info/detail*: find (specific) information/understand (specific) detail – find a specific piece of information or detail which is given in a text or picture.
2. *Main point*: understand the main point – identify the main point of a text or a section of a text.
3. *Interpret*: interpret and understand – interpret or have a more intuitive understanding of the text – the information required is not to be found directly in the text.
4. *Grammar*: understand/use grammatical structures – select or provide a particular grammatical structure (syntax/morphology/function word).
5. *Cohesion*: understand cohesion – put a series of disconnected paragraphs in a text into the right order.

The coding of these items was largely based on the wording in the task, although sometimes it was necessary to consider the text, e.g. to see if the answer was explicitly stated.

A number of statistical tests produced some evidence that, on both dimensions, certain features could be regarded, with varying degrees of significance, as predictors of difficulty in our test items. In the case of reading processing (Dimension 1), it was very evident that items requiring only an *understanding of words and phrases* were found to be easier than those requiring *understanding a sentence or clause*; a tendency was shown, moreover, that difficulty increases further when it is necessary to *make links between non-adjacent sentences* in a text.

With regard to Dimension 2, the operation required by the task, features were found to cluster into three levels of difficulty. At the simplest level was *finding information/detail*. Of intermediate difficulty were *understanding the main point* and *interpreting/inferencing*. The two most difficult operations were found to be *cohesion* (ordering paragraphs) and, at the extreme end, *grammar* (involving choosing a grammatical form).

It was also shown, unsurprisingly that length and readability index were significant predictors of difficulty. Taken together, these findings were heartening, as they largely reflect (or have later influenced) the progression in the descriptors in the five 'mastery' bands, used for reporting/interpreting test results (see Appendix 3).

4. Our testing group

Our testing group – currently nine of us in positions of various 'shapes and sizes' – have developed hand-in-hand with the tests we work on. The advice we received at the start has proved prophetic. We can confirm that digital tests are quite different from paper tests and need to be designed as such from the start. The restrictions imposed by digitalisation can be overcome with creative solutions, and are more than offset by their advantages, including the access to instant data from many thousands of test takers. The acquisition of a professional illustrator and a working primary school teacher as central team members has been invaluable. Our illustrator translates our ideas readily into backgrounds, objects and people to bring our tasks to life and give them great child-appeal. The teacher is able to use her knowledge of children and teachers, to inform us on what is feasible and beneficial. And she is able to talk to teachers in their own language, in the written guidelines and on courses we hold. We have also kept to the principle of having several native speakers on the team to ensure no *NorwEnglish* creeps into our tasks. Our statistician had helped us understand how our items work and has imparted much of his knowledge and skill to the group. He has advised us on mini projects and guided us through CEFR-standard setting of our tests.

As item writers we have, from the outset, held regular meetings where all our draft items are discussed by other team members and revised. A major innovation to our practice has been the coding of items from their conception, on the two dimensions outlined above. This has not only ensured that we have a deep common understanding of what the items test, but also that we maintain a balance of the different aspects, both in our item production and when test sets are put together.

We have no research funding but have passion for finding out more about our tests, and we regularly conduct small projects. In the pipeline are investigations into why some items 'just don't work' (in the piloting) and whether we can find characteristics of items that seem to discriminate well with 8th grade pupils but not with 5th graders.

In the fifteen or so years we have existed as a group we have learnt a lot but hope to learn still more. We have had many obstacles to overcome, particularly in the early years, when we were very much in need of Sauli's helping hand. We have his photo smiling down on us in our office, and we like to think he would still approve of what is going on!

References

Alderson, C. J. (2001). *Assessing Reading.* Cambridge: Cambridge University Press.

Cameron, L. (2001). *Teaching Languages to Young Learners*. Cambridge: Cambridge University Press.

Council of Europe (2001).*Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.

Hasselgreen, A., Grocott, C. & Torsheim, T. (2017). Quality Assurance in the National Tests of English: Investigating What Makes Reading Difficult. *Remaining Relevant. Modern*

Language Studies Today. Bergen: Bergen Language and Linguistics Studies, vol. 7, 216–233.

Hasselgreen, A. & Caudwell, G. (2016). *Assessing the Language of Young Learners*. Sheffield: Equinox Publishing.

Hasselgreen, A., Kaledaite, V., Maldonado Martin & N. Pizorn K. (2012). *Assessment of Young Learner Literacy Linked to the Common European Framework of Reference for Languages.* Strasbourg: Council of Europe.

Khalifa, H. & Weir, C.J. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*. *Studies in Language Testing 29.* Cambridge: Cambridge University Press.

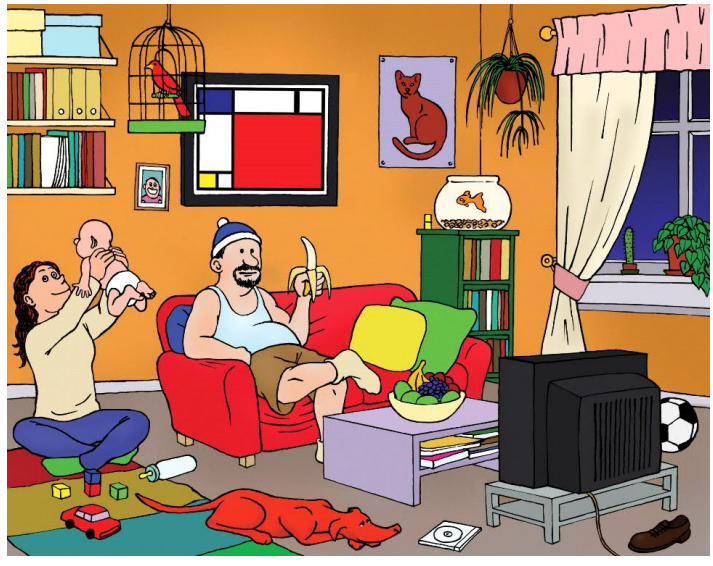
Nippold, M.A. (2007). *Later Language Development: School-Age Children, Adolescents, and Young Adults*. Austin, Texas: Pro-Ed.

Piaget, J. (1926). *The Language and Thought of the Child.* New York: Harcourt, Brace.

Appendix 1: Writing test and criteria - grade 7

Time: 60 minutes
Answer all three tasks. Spend half the time on task 3.

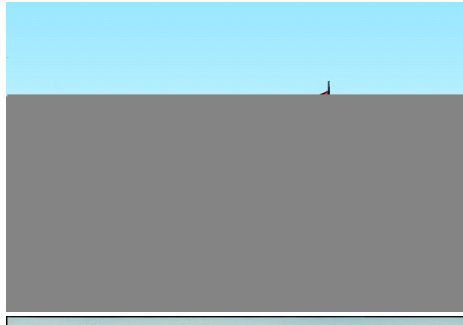
1. Look at the picture. What do you see?

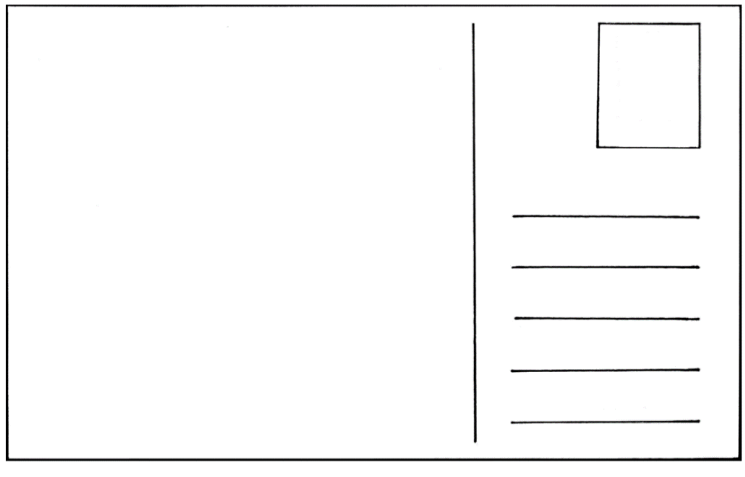


_____
_____
_____
_____
_____
_____
_____
_____

1. You are on holiday in one of these places. Write a postcard to a friend and tell him or her, for example:

   WHERE you are. WHAT you are doing. WHAT you like and/or dislike.





3. DO TASK a) OR b):

a) Write a text, 3 or 4 paragraphs. Start with the sentence:

One day when I came home from school, I found the front door wide open.

Give your text a title.

OR

b) Write your own text, 3 or 4 paragraphs, using this picture.

Write your text on the next page.

Assessment criteria 7th grade writing

Communication criteria:

7th GRADE: ASSESSMENT FORM FOR COMMUNICATION OF MEANING (TASK SPECIFIC) National tests of English						
Task 1: Description of a picture		**Task 2: Postcard**		**Task 3: Story/description**		**Level**
Can write a number of words, including some that are less common. Some extended phrases are also used, such as: *a fat bloke eating a banana, an lady with a baby,*		Can communicate a clear and coherent message in a friendly and informal tone, including some details.		Can write a generally clear and intelligible story or description. Can produce a quite logical and thematically coherent text.		B1
						A2/B1
		Can generally communicate a simple message in an understandable way, using a series of simple sentences.		Can write a short and simple story or description using a number of simple sentences.		A2
						A1/A2
Can write some of the most common words in an understandable way. Can write some simple fixed phrases, such as: *It's, I see, I can see.*		Can communicate parts of a message, using very simple words and phrases.		Can communicate parts of a message using very simple words and phrases.		A1
Communication of meaning:	**Language:**	**Total:**	**Comments:**			

Language criteria:

Textual structure		Grammar		Words and Phrases		Spelling and Punctuation		Level
Organisation: Thoughts and opinions are grouped into paragraphs to some extent. **Logical/thematic development:** On the whole, it is possible to follow the thread through the text. **Cohesion and flow:** There is some cohesion in the text and parts flow quite smoothly. A number of smallwords are used (e.g. *because, then*, etc.).		**Correctness:** Displays a relatively good grasp of the basic grammatical structures, although there may be frequent mistakes. **Complexity:** Some variation in sentence types, including some successful use of complex sentences. Some L1 influence is evident.		**Vocabulary:** The range of vocabulary is sufficient to communicate quite a lot on familiar topics, with some repetition. Words may be used imprecisely. **Idiomatic language:** Some idiomatic language is used correctly, although there may be considerable L1 influence. **Style/register:** Language use is adapted to context to some extent.		**Spelling:** Displays a relatively good grasp of English spelling, although spelling mistakes may be frequent. **Punctuation:** Punctuation is sufficient to indicate sentences. Commas and apostrophes are used occasionally.		B1
								A2/B1
Logical/thematic development: There may be some logic in the order thoughts are presented. **Cohesion and flow:** The texts consist of simple sentences, which are sometimes linked using smallwords (e.g. *and, so, but*, etc.).		**Correctness:** Displays some awareness of the basic grammatical structures. **Complexity:** Can construct simple sentences. The language may be strongly influenced by L1.		**Vocabulary:** Can use the most common words and phrases, which are sufficient to write about simple and everyday topics. L1 influence on the way things are expressed may be considerable, but it is generally possible to understand what is meant.		**Spelling:** It is usually possible to recognize the words, although mistakes occur in even the most common words. **Punctuation:** Some basic punctuation is present, including full stops and capitals.		A2
								A1/A2
The texts often consist of isolated words and phrases.		Can use certain fixed formulations correctly.		Can use some of the most common words and phrases, but there may be a lot of L1 expressions.		Can write some words correctly, which have been practiced or copied; for example, from the prompt.		A1

Appendix 2: Examples of item formats for National Test of English

Example 1 – Click item (for grade 5)

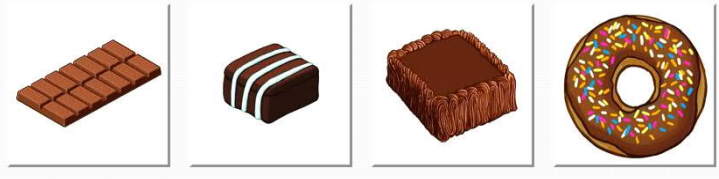
Read the text. Click on the correct person or item.

There are four backpacks in the classroom. Click on the one in front of the teacher.

Example 2 – Click picture (for grade 5)

Read the text. Click on the correct picture.

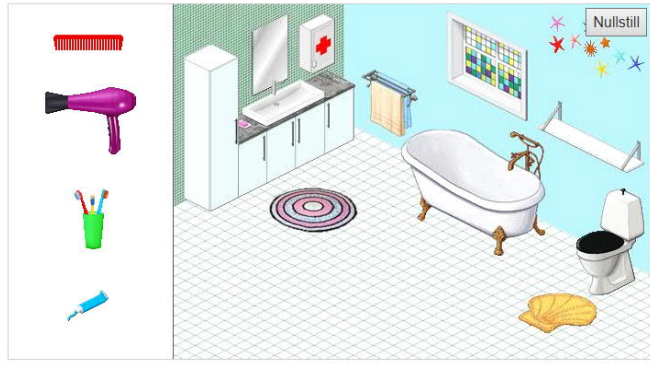
I love chocolate! I'll eat it in any form, but my all-time favourite is chocolate doughnuts with sprinkles.



Example 3 – Click and drag (for grade 5)

Read the text. Click and drag.

Liezel goes to brush her teeth but can't find her toothbrush anywhere. She finally finds all the toothbrushes in the bathtub. She suspects her two-year-old brother had something to do with this. Put the correct item where Liezel finds it.



Example 4 – Click word (for grade 8)

Read the text. Click on the correct word.

Click on the word that, in this context, means almost the same as 'answer'.

An airport received many complaints from passengers who felt they had to wait for too long for their baggage after arrival. After investigating the situation, the airport realised that it only took one minute to walk from the arrival gate to the baggage claim area. The solution was to move the arrival gate farther away. Making the passengers walk farther meant that they didn't have to wait so long and the complaints stopped.

Ditt svar: **solution**

Example 5 – Move paragraph (for grade 8)

Read the paragraphs below. Click on the arrows to put the paragraphs in the correct order. The first paragraph is given.

Linda and Anne are at a boarding school in the Scottish highlands. One day, they were peacefully skiing along a narrow path when they saw someone walking towards them.

This gave the girls a great idea! The next day they snuck into Ivan's room when he was in the cafeteria. They tiptoed over to his desk where the nearly completed puzzle was laid out. Anne grabbed one puzzle piece and put it in her pocket and the two girls snuck back out of the room without being noticed. Ned

They recognised the person as Ivan who lived in the same block as them. They stopped to let him pass when all of a sudden he pushed them, making them lose their balance and fall over in the snow. Ivan thought it was hilarious, but Linda and Anne weren't as amused! They told him that someday, when he least expected it, they would pay him back! Opp Ned

Ivan looked very sad and the girls started to feel a bit guilty. Before they went to bed they placed the puzzle piece in a little box with a note and gave it to Ivan. The note said, "Are you feeling puzzled? We hope that all the pieces fall into place. We got you!" Luckily Ivan took it with a smile and actually thought the prank was very clever. Opp Ned

Linda and Anne spent many hours discussing their revenge on Ivan. One evening Ivan showed them a puzzle he was working on. It was an impressive 1000-piece puzzle of a leopard. Ivan seemed very proud of his work and promised to show them the puzzle when he finished. Opp Ned

Later that day they heard strange sounds coming from Ivan's room. During dinner he told them that one piece was missing from his puzzle and that he had looked everywhere for it. He had searched high and low, under the bed and in the trash. He had even contacted the cleaning staff and asked them if they had seen it. Linda and Anne looked at each other and tried not to laugh. Opp

Appendix 3:1: Mastery level descriptors for reporting results on National Tests of English

(Levels 1-3: 5th grade, levels 1-5: 8th grade)

Mastery Level 1

Pupils

- Can understand some concrete, common words and expressions
- Can find common, concrete words in a text
- Can follow clear, simpl:e instructions
- Can link common, concrete words to pictures
- Can make links between familiar, concrete words within a theme, e.g. fish and acquarium
- Can recognise some learnt grammatical expressions and simple function words in context, e.g. personal pronouns.

Mastery Level 2

Pupils

- Can understand a number of common words and expressions
- Can understand simple sentences
- Can link simple sentences to pictures
- Can make links between common words in a text, when they are within a theme
- Can find specific details in a longer text
- Can find simple synonyms in a short text
- Can understand the main point in a simple text
- Can find simple information even when there is some competing in formation in a text
- Can navigate back and forth in a text to find information
- Can draw simple conclusions when there is a good deal of support in the text
- Can recognise and use some simple function words and grammatical structures in context.

Mastery Level 3

Pupils

- Can understand rather abstract and less common words and expressions
- Can construct meaning from some complex sentences
- Can construct meaning from shorter and longer texts
- Can understand the main point in a text
- Can find information even when there is competing information in a text
- Can read a text closely
- Can understand how the paragraphs in a text relate to each other
- Can link simple information from different parts of a text

Can use the context to understand difficult parts of a text
Can draw simple conclusions
Can recognise and use basic grammatical structures/ function words in context

Mastery Level 4
Pupils

Have a fairly wide vocabulary
Can work out the meaning of unknown words from the context
Can understand quite complex sentences
Can understand quite long and complex texts
Can link information from different parts of a text
Can draw conclusions
Can make choices between some grammatical structures/ function words in order to express him/herself.

Mastery Level 5
Pupils

Can use appropriate reading strategies
Have a quite wide and sophisticated vocabulary
Can understand complex sentences
Can understand long and complex texts
Can read between the lines and draw advanced conclusions
Can make choices between a range of grammatical structures/ function words in order to express him/herself

Finnish 9th graders' language skills: Effects of learning environment and teaching on levels attained compared with other European countries

Raili Hildén
University of Helsinki

Marita Härmälä
University of Jyväskylä

Juhani Rautopuro
University of Jyväskylä

Mari Huhtanen
Finnish Education Evaluation Centre

1. Introduction

Following the call of 21st century skills voiced by the OECD, European language policies have put increasing emphasis on the standardisation and mutual understanding of the objectives, contexts and outcomes of language teaching and learning. Among the most prominent tools not only in Europe but also globally is the Common European Framework of Reference or CEFR (Council of Europe 2001), which exercises great impact on national language curricula. The CEFR 6-level proficiency scale has, in many countries, been worked on into finer-grained national scales to allow for illustrating even small steps in foreign language acquisition. In Finland, Sauli Takala was actively participating in designing the CEFR-linked scales for the Finnish basic education (Hildén & Takala 2007).

The conceptual framework of language education in Europe is grounded in language use and learning as social activity (Vygotsky 1980). Therefore, a great emphasis is put on environmental factors framing this activity. These factors include for example regional factors, size of the school, school staff, support to students, school's leadership culture, and teacher-student relationships manifested through study practices. OECD as well as national European governments dedicate a lot of attention to the issues of equality between groups of students with regard to gender, parental socio-economic status and language groups (OECD, 2016). In language testing, the model of communicative language ability has remained relatively stable through

decades resorting to the four skills coined by Lado (1961), even if various imaginary contexts are provided in the task descriptions to simulate real-world use. The contexts, text types and themes are typically drawn on the national curricula.

The initiative of comparing achieved outcomes of English studies at the end of compulsory education was undertaken in 2011 when 15 European countries participated in the European Survey on Language Competences (ESLC) (European Commission, 2012). However, Finland was not one of these countries. Following the statement of the European Commission to contextualise the language tests by factors taking into account the circumstances of learning and teaching, we set out to chart the outcomes of language education at the national level and to compare them across other European countries. Our focus in this article is on 15/16-year-old students who are about to finish their compulsory education.

2. Context of the study

Compulsory basic education, which was introduced in Finland in the 1970s, consists of the lower (grades 1-6, ages 7-12) and the upper level (grades 7-9, ages 13-15). The most frequently studied language in the basic education is English (90 % of students) and it is usually started in grade 3 when the students are 9 years old. Studies of the other national language, Swedish (Finnish for the Swedish-speaking population), are mandatory and may be taken according to the aims of an advanced (started in grade 3) or short syllabus (started in grade 7). In grade four or five, one fourth of the students take an additional foreign language (Finnish National Agency of Education, 2014). Students may also choose another optional language in grade eight or nine but only about 11.5 % made this choice in 2016. The most frequently studied optional foreign languages in Finland are German, French, Spanish, and Russian.

During the last decades, there has been a considerable decrease in the Finnish students' language studies, optional languages are chosen increasingly seldom and the studies of new languages are of a short duration (Pyykkö, 2017). Due to this, new initiatives have been taken to promote language studies in the basic education. These initiatives include starting Swedish studies already in grade six, and the first foreign language (other than English) in the first grade (at the age of 7). Both initiatives aim at making the Finns´ skills in foreign languages more versatile.

By the end of basic education, in practice all students have studied English for seven years and Swedish for at least three years. The distribution of annual lessons in the advanced syllabus language is 16 (8+8), in the short syllabus six, and in the optional language 12 (6+6). One annual lesson equals 38 lesson hours of 45 minutes in duration. This means that at the age of 15, the students have studied the advanced English syllabus at least for 608 hours, the short Swedish syllabus for 228 hours, and one optional language for 456 hours. The target levels for good skills (grade 8 on a scale 4-10) are defined in the national core curriculum by using a Finnish application of the CEFR scales (Council of Europe, 2001; Hildén & Takala, 2007; Huhta, 2016). In these scales, the six CEFR proficiency levels have been divided into fine-grained sublevels,

that is, into A1.1, A1.2, A1.3, A2.1, B1.1, B.1.2 etc. The target level varies according to syllabi, e.g., for the advanced English syllabus it is B1.1 and for the advanced Swedish syllabus A2.2 (Finnish National Board of Education, 2004). In the current curricula, the target level of English is B1.1 in all four skills (Finnish National Agency of Education, 2014).

As there are no national tests in basic education, the learning outcomes are evaluated by an external evaluation body, that is, the Finnish Education Evaluation Centre (until 2014 by the National Agency for Education). In 2013, there was a national evaluation of learning outcomes in the most studied foreign languages in the Finnish basic education. This evaluation was carried out as a part of the national educational policy based on providing advice and support by information. Data gathering was sample-based and comprised 11 000 students from 661 schools. The students' competences were assessed in reading, listening, and writing. A smaller sample of students did the speaking tasks also. The languages assessed were English, Swedish, French, German, and Russian. The results were reported by using the fine-grained Finnish version of the CEFR proficiency scales. Another part of the data gathered consisted of questionnaires, which were answered by the students, teachers, and school principals.

In this article we compare the results of the advanced English syllabus and advanced Swedish secondary syllabus with the results of the ESLC (European Commission, 2012). These languages were chosen because of their central role in the Finnish language syllabus. The closest counterpart for the European data for "second language" would have been the intermediate course of Swedish, which is studied as a mandatory syllabus at the upper grades of basic education by all the pupils who have not started it earlier. The learning outcomes of that syllabus have not been evaluated since 2007, which makes the data outdated for our purpose. Therefore, we chose the advanced Swedish secondary syllabus, which starts in grade 5 of basic education. Swedish is the most popular language studied at that stage (next to English, 8.3%) selected annually by 7.2% of the students.

We start by describing the participants and the tasks used in the two tests. We then present the results in both languages by subskills and also as a composite score. To conclude, we discuss the differences and similarities between Finland and the other countries, in particular with Sweden and Estonia, as well as propose measures to enhance language teaching and learning. Sweden is chosen for comparison for the similarity of societal and educational systems dating back to the countries' common history until 1809. On the other hand, Finnish and Estonian languages are linguistic relatives, and their educational systems resemble each other today.

3. Research questions

The research questions are:

1) What is the percentage of Finnish students at each CEFR level using the global average of the 3 skills compared with the other European countries?

2) What is the relationship between Finnish students' language proficiency and informal language learning, teaching methods and curricula compared with the other European countries?

To answer the first question, we counted a composite score for the three skills (reading and listening comprehension, writing) which we then compared with the European levels. For question two, we chose in the background questionnaire questions, which were related to the use of media and study practices both in school and on free time and consequently allow comparisons with the ESLC results. Consequently, we are to some extent able to compare the results of the Finnish study with those obtained in the European study. At the same time, we are fully aware that these comparisons are only tentative as for example the meter and the methods used in the two studies are not the same.

More detailed results of the Finnish 2013 national study can be found in its main reports (Härmälä, Huhtanen & Puukko, 2014; Hildén & Rautopuro, 2014) as well as in an article published by Härmälä, Leontjev and Kangasvieri (2017) where the relationship between students' opinions, background factors, and learning outcomes in English was explored and modelled.

4. Data

The data are twofold. First, they consist of the results of the tests in reading and listening comprehension and in writing. Second, the students and teachers answered a background questionnaire inquiring on their perceptions of studying the language in question and on their study practices during the language lessons and in free time. The teachers were also asked what teaching contents, e.g. grammar, vocabulary, speaking, they considered important when giving their students the final marks.

The number of students in the main study is summarised in table 1. In the tests of English, there were students from both Finnish and Swedish-speaking schools.

Table 1. Number of students and teachers per language in the Finnish data.

	English	**Swedish**
Number of students	3 476 (Fin 2 966 + Swe 510)	1 487
Number of teachers	220	81

In both languages, the test tasks and items were written by language teachers and experts in language testing. The item types in the receptive skills were multiple choice and open-ended questions. In the Finnish study, all the instructions in the test booklet were given in the students' L1. In writing, two tasks were used: one shorter (40-60 words) and the other a bit longer (80-150 words). The tests were administered in paper and pencil format and performed under teacher supervision. The total time for the test administration was 130-140 minutes. After the tests, the teachers assessed the answers with the help of an answer key (OE items) and benchmarks performances (writing). Afterwards, 10 % of the performances were rated at the National Agency for Education. To set the standards between different proficiency levels, the Bookmark method was used (Cizek, 2011).

In appendixes 1 and 2, there is a summary of the tasks in both languages. The summary includes the themes, text types, levels, item types and the number of items in reading, listening and writing (see also Härmälä, Huhtanen & Puukko, 2014; Hildén & Rautopuro, 2014).

In the students' background questionnaire, three items were chosen for comparison with the European data. The items enquired on what the students did in English/Swedish lessons:

17. We watch films / listen to songs etc. in English.
22. We practise grammar on computer (games etc.).
23. We use the internet to search for information.

These items had a counterpart in the teacher questionnaire verbalised as "In my language class, the pupils watch video films…" and so forth. The temporal 5-point scale ranged from "Never" to "Almost in every lesson" for both teachers and students. In the student questionnaire, there were also eight items enquiring on the students' extra mural use of the target language:

33. I watch films or video clips in English.
34. I listen to music in English.
35. I follow discussion forums in English (Youtube, Facebook etc.).
36. I participate in web discussions in English (blogs, Facebook, chat, Twitter).
37. I read magazines and other texts in English in the internet.
38. I write texts in English (sms, poems etc.).
39. I speak English e.g. with tourists.
40. I use English with my friends or relatives.

The 5-point scale for these items was from "Never" to "Every day" for both the teachers and the students.

On the basis of the student questionnaire (SQ), three sum variables (means of several items) were constructed:

1) Use of the media (items 33-38 above)
2) Use of the ICT (items 17, 22, 23 above)
3) Use of the target language (items 39 and 40)

The reliability coefficients of these scales are presented in Table 4.

In the teacher questionnaire, the following sum variables were aggregated to reflect environmental factors of language learning:

- Environmental responsibility: Among curricular objectives the teacher values Human and technology, Safety and traffic, Responsibility for the environment, welfare and sustainable futures.
- Written language use: Among curricular objectives the teacher values writing skills, grammar and reading.
- Cultural agency: Among curricular objectives the teacher values a student's growth as a person, encountering diversity, developing cultural identity and internationalism, participatory citizenship and entrepreneurship, Media skills and communication.
- Cultural communication and oral language use: Among curricular objectives the teacher values Vocabulary, Courage of expression, study skills, cultural knowledge, communication strategies, spoken interaction, spoken production.

The scale used for prioritising the objectives perceived by teachers ranged from 1 (not important at all) to 5 (highly important).

5. Analysis and results

In the following we compare the learning outcomes of Finnish students at the end of compulsory basic education with the overall results of the European survey, and more specifically the language proficiency attained in the neighbouring countries Sweden and Estonia.

The scales described earlier, and a number of items from the student and the teacher questionnaires were used to predict the level of students' language skills. Linear regression analysis was applied to model the associations between the variables. For counting the composite indicator, we used the mean value of the percentages of right answers. We present the results by answering the two research questions.

5.1 The Finnish students results in reading, listening and writing and as a composite indicator across skills compared with other European countries

First, we take a look at the results of the first foreign language, English, by subskills. For the receptive skills, we were able to distinguish between four levels (A2.1 or below, A2.2, B1.1, and B1.2 or higher), that is, two CEFR levels (A2 and B1) due to the small number of items in each subtest.

In **listening**, 33% of the Finnish students displayed proficiency at the levels A1-A2, the corresponding figure for the EU countries participating in the ESLC was 52%. (level pre-A1 included) (ESLC 2011: 91). For level B1, the percentages were 67 % (Fin) and 16 % (EU). The level B2 was attained by a third (32%) of European students, in Finland the 2013 study could not identify the level B2 for the reasons mentioned above. 48% of European students against 67% of Finnish ones were assigned at B-levels. In sum, the Finnish students achieved on average higher in listening than their European counterparts. (Härmälä, Huhtanen & Puukko, 2014: 71, ESLC 2011:35, 91)

When comparing the Finnish results in listening with Sweden and Estonia, the level B1 or higher was attained by 91% of Swedish and 37% of Estonian students (ESLC ,2011: 92). The percentage of lower achievers on levels A1-A2 was only 10% in Sweden and 37% in Estonia. In sum, the Finnish students exceeded the European average and Estonian students in listening, but did not attain the level of Swedish students. (Härmälä, Huhtanen & Puukko, 2014: 71; ESLC, 2011:35, 92)

In **reading**, 38% of Finnish 9th graders were placed in levels A1-A2, while the European average was 58% (level pre-A1 included). Consequently, the level B1 or higher was achieved by 62% of the Finnish and on average by 42% of the European students (Härmälä, Huhtanen & Puukko, 2014: 71; ESLC 2011:35). In Sweden, the figures were 19% vs. 81%, and in Estonia 40% vs. 60%. In reading, the Estonian and the Finnish students performed quite equally, while Sweden held the lead throughout. (Härmälä, Huhtanen & Puukko, 2014: 71; ESLC, 2011: 92)

The only productive skill measured in both studies was **writing**. In writing, a more fine-grained comparison was possible, as the performances were rated by assigning them directly to the CEFR levels. The European average for A-levels was 57% (pre-A1 included), but only 43% of the Finnish students remained at these levels. The percentage of the pre-A1 was on average 9%, in Finland 4%. 27% of the Finnish students reached level A2, while the European average placed in this level was 24%. With regard to the level B1, 29% of the European and 39% of the Finnish pupils attained it in writing. Level B2 writers comprise 14% of the European and 19% of the Finnish students. (Härmälä, Huhtanen & Puukko, 2014: 71; ESLC, 2011:35)

In Sweden, less than 1% of the students were placed under A1 in writing, 25% attained levels A1-A2, and 75% were assigned at the levels B1-B2 (against 57% of the Finnish students at these levels). In Estonia, 3% remained at the pre-A1 level and 60% attained the B-levels. (Härmälä, Huhtanen & Puukko, 2014: 7; ESLC, 2011: 92) Unlike

in receptive skills, the Estonian students clearly outperformed the Finnish students in writing. Reasons may be found in the different task demands or the different traditions of language teaching between the countries. The excellence of the Swedish students is partly due to the fact that Swedish and English are rather close linguistic relatives as opposed to the Finno-ugric language group that both Finnish and Estonian belong to. Support for this assumption is provided by the finding that in the Finnish data, the students of the Swedish-speaking schools outperformed those from the Finnish-speaking schools. The summary of the results in both languages is presented in tables 2 and 3. Table 2 summarises the CEFR levels attained by the Finnish students.

Table 2. The CEFR levels attained by the students in advanced syllabus English in Finland.

Advanced syllabus English					
Skill	Beginner Pre-A1 (Fi A1.1)	Basic A1 (Fi A1.2-A1.3)	Advanced Basic A2 (Fi A2.1-A2.2)	Independent B1	Advanced Independent B2
Listening			33*	67*	
Reading			38*	62*	
Writing	5	11	27	39	19

*These figures include also the preceding / following levels.

In the ESLC, a "composite" indicator for the students' language proficiency was produced by averaging across language skills. This was done by taking the average of the proportion of students achieving each CEFR level in reading, listening and writing. The composite indicator counted in this manner resulted in the following distribution: 38% of the students achieved A2 (levels A2.1. and A2.2 combined) and 62% B1.

The results for the second target language (second advanced Swedish syllabus in the Finnish data), are discussed next. The most typical second languages in the ESCL data were French, German and Spanish, but the provision of these languages in Finland is clearly lower than the second national language Swedish studied as a mandatory syllabus in the general education.

In **listening**, 85 % (pre-A1 included) of the European pupils attained levels A1-B1 (pre-A1 included) in their second target language, while 78% of the Finnish students learning second advanced syllabus Swedish were placed at levels A1-B1.1. The levels of an independent user, B1-B2, were reached by 29% of the European students, 22% of the Finnish students were placed at levels B1.2 or above. In **reading**, 84% EU students reached the levels A1-B1, and 78% of the Finnish students in second advanced Swedish syllabus reached the levels A1-B1.1. In **writing**, 77 % of the European students and 81% of the Finnish students were placed at levels A1-A2. The levels of an independent user B1-B2 were attained by 23% of the European students and by 19%

of the Finnish students learning the second advanced Swedish syllabus . Table 3 summarises the levels attained in the second advanced Swedish syllabus. (ESLC, 2011: 35, Hilden & Rautopuro, 2014: 72.)

Table 3. The CEFR levels attained by the students in the second advanced Swedish syllabus in Finland.

Second advanced Swedish syllabus					
Skill	Beginner Pre-A1 (Fi A1.1)	Basic A1 (Fi A1.2-A1.3)	Advanced Basic A2 (Fi A2.1-A2.2)	Independent B1	Advanced Independent B2
Listening			50*	50*	
Reading			52*	48*	
Writing	11	33	38	15	3

*These figures include the preceding / following levels.

Next, we answer our second research question and model the relationships between the variables analysed.

5.2 Relationship between language proficiency and informal language learning, teaching methods and curricula compared with other European countries

The reliability coefficients (Cronbach's alpha) of the scales measuring correlations between the use of media and the language proficiency were high in both languages for the use of media (alpha 0,81), moderate for the use of ICT (0,60) and for the use of the target language in free time (0,64). Descriptive statistics of these scales are presented in Table 4.

Table 4. Descriptive statistics and reliability of the scales between the use of media and language proficiency in English and Swedish in Finland.

	Mean ENG / SWE Scale 1-5	Std. Deviation ENG / SWE	Alpha ENG / SWE
Homework	4.1 /4.1	1,0 /1,1	--
Use of Media (SQ, 6 items)	2,9 / 1,4	0,7 /0,4	0,81/0,82
Use of ICT (SQ, 3 items)	2,3 /2,2	0,7 /0,7	0,60/0,63
Use of TL in free time (SQ, 2 items)	2,0 /1,5	0,7 /0,4	0,62/0,65
Usefulness (SQ, 5 items)	4,2 /3,5	0,7 /1,0	0,82/0,88
Liking (SQ, 5 items)	3,4 /2,6	1,0 /1,0	0,87/0,90
Teacher speaks TL (SQ)	3,9 /3,8	1,1 /1,0	--
Teacher uses TL (TQ)	4,1 /4,0	0,7 /0,7	--
Environment (TQ, 3 items)	3,1 /3,2	0,9 /0,7	0,87/0,85
Written language use (TQ, 3 items)	4,2 /4,1	0,5 /0,5	0,77/0,75
Cultural agency (TQ, 5 items)	3,7 /3,9	0,7 /0,6	0,84/0,84
Cultural communication and oral language use (TQ, 6 items)	4,1 /4,1	0,4 /0,5	0,66/0,79

As Table 4 clearly demonstrates, the highest averages in both Finnish datasets were assigned to students doing their homework, to teacher using TL (reported by students), teacher's self-reported use of TL, to valuing written language use and to cultural communication and to the teachers use of oral language. All the activities embedded in these variables were carried out often or very often in both English and Swedish classes. Differences between the two syllabi were detected with regard to the usefulness and liking the language as a school subject. English was perceived as being more useful than Swedish and, on average, the students also liked to study English more than Swedish. Both findings date back to the current linguistic reality in Finland supported by the media's massive use of English. On the other hand, Swedish, despite its official status as the second domestic language in Finland, is rarely used on a voluntary basis in freetime in areas other than in bilingual regions along the western coast.

In the Swedish data in Finland, the best predictors of good performance in all subskills were doing homework regularly, using the target language in free time, the perceived usefulness of Swedish and liking it, as well as the emphasis the teacher put on the written language use in instruction. The impact of doing homework emphasises

the nature of language study as a long-term endeavour. The use of Swedish in free time and its perceived usefulness likely refer to a bilingual environment, while liking the subject is more directly related to the school instruction. The sum variable of written language use comprises teacher appreciation of writing skills, grammar and reading in teaching. The effect of other variables proved to be more inconsistent and address only some subskills. The use of ICT improved reading, the teacher's use of Swedish in class improved listening, reading and the composite score. The cultural communication had an effect on reading and the composite score, but not on listening, although the sum variable distinctively includes the use of spoken language. The standardised regression coefficients for Swedish are summarised in table 5.

Table 5. Standardised regression coefficients between certain learning-related variables and the sub-skills of second advanced syllabus in Swedish in Finland.

	Subskills			
Independent variables	Listening	Reading	Writing	Composite score
Homework	0,14***	0,21***	0,23***	0,22***
Use of ICT	NS	0,07**	NS	NS
Use of media	0,07*	NS	NS	NS
Use TL in free time	0,11***	0,07**	0,15***	0,13***
Usefulness	0,17***	0,16***	0,18***	0,19***
Liking	0,10**	0,12***	0,13***	0,14***
Teacher speaks TL	NS	NS	NS	NS
Teacher uses TL	0,09**	0,06*	NS	0,07**
Environment	NS	NS	NS	NS
Written language use	0,14***	0,05*	0,10**	0,11***
Cultural agency	NS	NS	NS	NS
Cultural communication and oral language use	NS	0,12***	NS	0,06*
R-square	22 %	23 %	28 %	30 %

* p< 0,05 **p<0,01 ***p<0,001

In the Swedish data, the significant predictors of all the measured subskills as well as the composite score turned out to be homework, the use of the target language in free time, the perceived usefulness and students' liking the TL, and the written language use valued by the teachers. A somewhat more varying influence was detected for the use of ICT on reading and the media on listening, for the teacher-reported use of the target language on other scores but writing, and the effect of cultural communication and oral language use on reading.

In the English data, the strongest predictors were the perceived usefulness and liking of the TL as well as the role of English-medium media and the teachers' use of the TL in lessons. Surprisingly, the use of ICT in the English data showed a negative correlation with all the subskills. This finding corresponds fairly well with the European data where similar negative effects were found.

Table 6. Standardised regression coefficients between certain learning-related variables and the subskills of advanced syllabus English in Finland.

	Subskills			
Independent variables	Listening	Reading	Writing	Composite score
Homework	0,03*	0,08***	0,09***	0,09***
Use of ICT	-0,07***	-0,08***	-0,08**	-0,08***
Use of media	0,25***	0,28***	0,26***	0,30***
Use of TL in free time	NS	-0,05**	NS	NS
Usefulness	0,13***	0,18***	0,18***	0,19***
Liking	0,18***	0,16***	0,23***	0,21***
Teacher speaks TL	0,04**	0,05**	0,06***	0,06**
Teacher uses TL	0,07***	0,08***	0,08***	0,08***
Environment	NS	NS	NS	NS
Written language use	NS	-0,04*	-0,04*	-0.05**
Cultural agency	NS	NS	NS	NS
Cultural communication	NS	NS	NS	NS
R-square	23 %	26 %	34 %	35 %

* p< 0,05 **p<0,01 ***p<0,001

What were then the effects of the various factors on the composite scores in the three datasets? Table 7 presents the sum variables on the composite scores in the European data and the two Finnish datasets. Two sum variables (Environment and Cultural agency), which only yielded non-significant associations in all datasets, are left out from the table.

In the European data clearly positive effects were detected for the perceived usefulness and liking the TL and the use of it on free time, as well as for the teacher using the TL in class. The use of ICT had an unexpectedly negative association with language proficiency. In the Finnish data, the effects varied between languages. The effect of regular homework was positive in English and Swedish, so were the effects of perceived usefulness and liking, as well as the teacher-reported use of the target language in class. However, some effects were non-significant for one language only: the use of ICT had a small negative effect on the Finnish students' English proficiency, while the use of media was positively associated with it. The effect of using ICT and media were the most obscure sum variables with controversial effects on the proficiency in different languages. The finding deserves further exploration. The use of written language with an emphasis on grammar was quite naturally a slightly positive predictor of grammar and writing of first target language in the EU data and the Finnish data of Swedish. Somewhat unexpectedly, its effect was negative in the Finnish data of English – a finding that may be explained by the frequent occurance of oral and oral-like language use in social and other media.

Table 7. Effects of various variables on composite scored in the three datasets.

	EU	EN/FI	SWE/FI
Homework	mixed effects	0,09***	0,22***
Use of Media (SQ, 6 items)	slightly positive	0,30***	NS
Use of ICT (SQ,3 items)	slightly negative	-0,08***	NS
Use of target language in free time (SQ, 2 items)	positive	NS	0,13***
Usefulness (SQ, 5 items)	positive	0,19***	0,19***
Liking (SQ,5 items)	positive	0,21***	0,14***
Teacher speaks TL (SQ)	positive	0,06**	NS
Teacher uses TL	positive	0,08***	0,07**
Written language use	slightly positive for grammar and writing	-0.55***	0,11***
Cultural communication and oral language use		NS	0,06*

6. Discussion

In the ESLC (European Commission, 2012), the overall level of language competence in both first and second foreign language was found to be low. In Finland, however, the level of independent user (B1-B2), was achieved by the majority of the students (67%, 62%, 57%) in the first foreign language, and by around half of the students in the second foreign language (listening and reading). Consequently, the Finnish results were better in the first foreign language, which confirms the results of the European study. The better results in Finland can also be explained by the students' perceptions of English: it is a useful language and the students like to study it and also use it outside school. What distinguishes Finland from the other European countries is that also in the second foreign language the Finnish results were fairly good in listening and reading comprehension. In writing, however, only 16 % of the Finnish students achieved B1-B2 level.

When it comes to the exposure to foreign languages in Finland, the possibilities for both English and Swedish are extremely good at least in the southern parts of the country. All television programmes are broadcast in their original language, which facilitates early exposure especially to English, and also to some extent to Swedish, the other national language. The growing number of migrants has also increased the

linguistic and cultural diversity of the country, at the same time reducing the difference between second and foreign languages.

The European goal of 2+1 languages is therefore well met in Finland even though linguistic diversity remains a challenge as the Finns' language skills continue to narrow. It is, therefore, important that studies like ESCL are conducted regularly across educational systems. This, in turn, allows promoting the increasing multilingualism and linguistic diversity in Europe.

The comparison also revealed a few shared challenges in the EU countries. The European Language Policy Division has produced excellent tools for mobility, common understanding of the attained language proficiency levels and also tools for reporting it. Unfortunately, the European Language Portfolio, for instance, is scarcely used and language teachers do not profit from all the available opportunities of intercultural exchange at school level (Council of Europe, 2008). A fresh start to the potential use of the ELP is given by the new curricula putting additional weight on formative assessment and scaffolding to language learners from an early age.

To be able to encounter diversity, inside Europe and beyond, has grown in importance along with the increased migration and the introduction of new languages and cultures in all European countries. It is wise to maintain the strengths and best practices developed to date and apply and elaborate them for implementation of sustainable language policy and local pedagogies for peace and understanding.

References

Cizek, G. (2011). *Setting Performance Standards: Foundations, Methods, and Innovations* (2. edition). New York: Routledge.

Council of Europe. (2001). *Common European Framework of Reference for Languages.* Cambridge: Cambridge University Press. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Council of Europe. (2008). *European Language Portfolio*. Retrieved from http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/introduction.html

European Commission. (2012). *First European Survey on Language Competences. Final Report*. Retrieved from https://www.researchgate.net/publication/262877352_First_European_Survey_on_Language_Competences_Final_Report

Finnish National Board of Education. (2004). *National Core Curriculum for Basic Education.* Helsinki: Finnish National Board of Education.

Finnish National Agency of Education. (2014). *National Core Curriculum for Basic Education*. Helsinki: Finnish National Agency of Education.

Hildén, R. & Rautopuro, J. (2014). *Ruotsin kielen A-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013*. [National evaluation of learning outcomes in Swedish language at the end of compulsory basic education 2013] ISSN 2342-4176; 2014:1. 220 p. Helsinki: Finnish Education Evaluation Centre. http://karvi.fi/publication/ruotsin-kielen-oppimaaran-oppimistulokset-perusopetuksen-paattovaiheessa-2013

Hildén, R. & Takala, S. (2007). Relating descriptors of the Finnish school scale to the CEF overall scales for communicative activities. Teoksessa A. Koskensalo, J. Smeds, P. Kaikkonen & V. Kohonen (Eds.) *Foreign languages and multicultural perspectives in the European context*, pp. 291–300. Reihe: Dichtung – Wahrheit – Sprache Bd. 7. Münster: LIT Verlag.

Huhta, A. (2016). Using the Common European Framework of Reference in the evaluation of educational achievement in foreign and second languages. In *Proceedings of the Second International Conference for Assessment & Evaluation: 'Learning Outcomes Assessment'*, Riyadh, Saudi Arabia, 1-3. December 2015. pp. 324–342. Available at http://ica.qiyas.sa/downloads/Conference_Book.zip

Härmälä, M., Huhtanen M. & Puukko M. (2014) *Englannin kielen A-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013* [Learning outcomes in English at the end of basic education 2013]. Helsinki: National Education Evaluation Centre. http://www.oph.fi/download/160066_englannin_kielen_a_oppimaaran_oppimistulokset_perusopetuksen_paattovaiheessa.pdf

Härmälä, M., Leontjev, D. & Kangasvieri, T. (2017). Relationship between students' opinions, background factors and learning outcomes: Finnish 9th graders learning English. *International Journal of Applied Linguistics, 27*(3), pp.665-681.

Lado, R. (1961). *Language Testing. The construction and use of foreign language tests.* London: Longmans.

OECD (2016). *PISA 2015 Results (Volume 1): Excellence and equity in education.* Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264266490-en

Pyykkö, R. (2017). *Monikielisyys vahvuudeksi. Selvitys Suomen kielivarannon tilasta ja tasosta.* [Multilingualism into a strength. A report of the status and levels of language competences in Finland] r Publications of the Ministry of Education and Culture, Finland 2017:51 http://urn.fi/URN:ISBN:978-952-263-535-8

Vygotsky, L.S. (1980). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

APPENDIX 1: Task examples

Tasks in the listening comprehension (English)

Title/ name	Theme	Text type	CEFR Level	Item type/ max. points
1. Rules of the game	freetime, sport	descriptive	A1.3–A2.2	3 mc / 3 p.
2. Ayrton Senna *	car sports, media	narrative, descriptive	B1.1	3 mc / 3 p.
3. Bonfire Night	culture	descriptive, expository	B2.1	3 mc / 3 p.
4. Discussions *	pets,travel	instructive	A2.1–A2.2	3 mc / 3 p.
5. Announcements	travel, food	instructive	A1.3–A2.2	3 oe / 4 p.
6. Weather forecast	weather	narrative	A2.2–B1.1	3 oe / 6 p.
7. Weekend tips	travel, culture	instructive	A1.2–B1.1	3 oe / 6 p.
8. Animal rescue	health, holidays	narrative	B1.1–B1.2	3 oe / 4 p.

* Listened to only once, all the others twice

Tasks in the reading comprehension (English)

Title / Name	Theme	Text type	CEFR level (Bookmark)	Item type/ max. p.
1. Presentation of Wales	travel, culture	descriptive	A2.1	3 mc / 3 p.
1. Clothing tips for boys	clothes, recy-cling	instructive	B1.1	3 mc / 3 p.
2. Interview with Cowell	work, media, music	descriptive, narrative	B1.2	3 mc / 3 p.
3. Short texts on recycling	sustainable development	instructive	A2.2, B1.2	3 mc /3 p.
4. News on kangaroo	work, living in country and city	narrative	A2.2-B1.1	3 oe / 5 p.
5. Letter from the Mordock-Bowers family	freetime, sports, family	narrative	B1.1	3. oe / 3 p.
6. News on a road accident	trafic, health, well-being	narrative	A2.2-B1.1	3 oe / 5 p.
8. Tips for a job interview	work	instructive	B1.1	3 oe / 3 p.

The writing tasks in English

Title / Name	Theme	Text type	Words	Planned level
1. Message to hotel reception	travel, public services	descriptive	40–60	A2.1–B1.2
2. Favorite book or film	freetime, films, litera-ture	narrative	80–150	A2.2–B1.2

Understanding self-assessment – what factors might underlie learners' views of their foreign language skills?

Ari Huhta
University of Jyväskylä

1. Introduction

Sauli Takala had a profound influence on my career in language assessment. His course on assessing writing in the Department of Applied Linguistics at the University of Jyväskylä in 1986 was the first course dedicated to assessment that I ever attended and also the first time I met Sauli. It was a unique course in many ways. First, it was based on cutting-edge knowledge about writing assessments as Sauli had been involved in coordinating the IEA Written Composition Study (see, e.g., Gorman, Purves & Degenhart, 1988) and had just returned to Finland from the USA. Secondly, he introduced us students to something completely new at the time: word processing by using a programme called WordStar. We needed to learn word processing because the outcome of the course was a series of chapters on learning, teaching and assessing writing, to be published in a publication series targeting language teachers in Finland. The course resulted in my first real scientific publication (Huhta, 1987).

In the following years, Sauli Takala was instrumental in planning and leading several projects that entailed developing new language assessment systems, conducting research on those systems, and organising training on assessment for teachers. I was involved in many of those projects and benefited greatly from Sauli's insights and expertise. One of the most significant projects was the creation of the Finnish National Certificates in the early 1990s as a joint operation between the University of Jyväskylä and the Finnish National Agency for Education (see https://www.oph.fi/ english/ services/yki). The National Certificates is a language proficiency examination system intended for adults who want to have their language proficiency certified for work or study purposes. Inaugurated in 1994, the examination now has nine languages and over 10,000 test takers each year, most of whom take Finnish as a second language examination.

Although Sauli was heavily involved in creating new high-stakes language examinations such as the National Certificates and in improving existing examinations like the Finnish Matriculation Examination (see Juurakko-Paavola's account in this volume), he was also very active in promoting assessment for, rather than of, learning and teaching. He was an advocate of portfolio assessment since the early 1990s, when

the first seeds for the European Language Portfolio were sown (Little, Goullier & Hughes, 2011) and discussed the benefits of the portfolio in several publications intended for researchers, teacher trainers and teachers (e.g., Linnakylä, Pollari and Takala 1994; Takala 1992, 1995). In his view, the portfolio integrates learning, teaching and assessment in ideal ways and, thus, improves learners' agency and understanding of the entire learning process. An important benefit of the portfolio, he argued, is that it improves learners' awareness of what they can do since self-assessment is a key aspect of the portfolio (see, e.g., the European Language Portfolio; Little, 2005; Little & Erickson, 2015).

Besides his work on promoting portfolio assessment, Sauli Takala was also involved in other approaches to assessment that were designed specifically to support language learning. The most notable of these was the DIALANG project (see Huhta, Luoma, Oscarson, Sajavaara, Takala & Teasdale 2002; Alderson 2005). A key component of that diagnostic on-line assessment and feedback system is self-assessment: DIALANG does not only contain language tests but also calibrated self-assessment instruments that the users of the system can take and receive feedback that relates to self-assessment.

Self-assessment as part of more extensive approaches to diagnostic or formative assessment was, thus, one of the many areas of applied linguistics that were close to Sauli's heart. Since I myself have been very interested in self-assessment, particularly as a consequence of my involvement in the DIALANG project, it is befitting to focus on this form of assessment in my contribution to the memorial publication. The roots of the study I will be reporting here go back to DIALANG but the empirical data for it were collected several years later in a very different type of study.

2. Self-assessment in DIALANG

DIALANG is an on-line diagnostic assessment system that provides its users with feedback about the strengths and weaknesses in their language proficiency in 14 different languages and five skill areas (reading, listening, writing, vocabulary, and structures). Besides language tests, the system includes self-assessment instruments for reading, listening and writing; replying to the self-assessment task is optional but recommended. Self-assessment in DIALANG is always related to the skill to be tested: thus, if the learner wants to take a test of reading – in any of the available test languages – they are given the opportunity to self-assess their reading in that language.

The self-assessment instrument comprises 18 statements that describe specific activities related to the skill in question, for example, reading based on the CEFR (for the design and validation of the self-assessment instrument, see Alderson, 2005). Users can choose the language in which to read the self-assessment statements from a list of 18 different languages since all self-assessment instruments have been translated from the original English into the 17 other languages. This enables as many users as possible

to read the statements in a language that they know well enough to conduct meaningful self-assessment. For example, a French-speaking learner of Spanish who is not very fluent in Spanish can study the statements in French before taking a test of Spanish. The users are invited to state whether they can do (or not) the activities described in each statement. The following examples illustrate some of the English-language versions of the self-assessment statements for reading:

(1) *I can follow short, simple written instructions, especially if they contain pictures.*
(2) *I can read correspondence relating to my fields of interest and easily understand the essential meaning.*
(3) *I can understand a wide range of long and complex texts, understanding fully subtleties of style and meaning which is directly stated or implied.*

Each statement has been linked to a specific CEFR (Common European Framework of Reference; Council of Europe, 2001) level. The above samples 1, 2 and 3 illustrate level A1, B1 and C2, respectively. The system treats the users' responses as if they had taken an 18-item language test covering CEFR levels A1 – C2 and calculates an estimate of their (self-assessed) CEFR level from the response data. After taking a test in the same skill, the learner is provided with feedback that reports which CEFR level most closely corresponds their self-assessment and which CEFR level they achieved in the language test. Thus, the users can see whether their self-assessed proficiency level matches the level they are assigned based on their test score.

The main function of self-assessment in DIALANG is to provide language learners with an opportunity to practice self-evaluation and improve their metacognitive skills and awareness of themselves as language learners through interacting with the self-assessment task and the related feedback. A secondary function of self-assessment is to direct the users to the most appropriate test version in terms of difficulty: based on learners' self-assessment and a vocabulary size test, the system administers them a basic, intermediate or advanced version of the test in the skill and language they chose (e.g., intermediate test of reading in Spanish).

DIALANG includes a complementary aspect of self-assessment feedback and it is this type of feedback that inspired the study reported here. Besides reporting on the match, or lack of it, between self-assessment and test result, as described above, DIALANG presents its users with an opportunity to read about potential reasons for a mismatch between self-assessments and tests. This part of feedback is titled *About self-assessment*, and its main screen is shown in Figure 1. By clicking on the links shown on the screen, learners can access more detailed information about the potential causes for the misalignment of the two.

This part of DIALANG feedback was somewhat difficult to design as we could not draw on any existing models for such feedback and relatively little research existed that could inform its content. Therefore, this feedback remained rather speculative. However, we felt that it would be important to give the users of the system

an opportunity to engage somewhat more deeply in thinking about their language skills, how they have acquired them, and how they use them. Therefore, this feedback focuses on increasing learners' awareness of their language skills, language learning, and metacognitive skills. Indeed, all feedback and information related to self-assessment in DIALANG illustrates what Hattie and Timperley (2007) refer to as self-regulation feedback that specifically focuses on improving learners' metacognitive skills. In addition to increasing learners' self-reflection with self-assessment, we also wanted to promote the value of self-assessment as an important and relevant approach to assessment in its own right and to counter the probably quite common assumption that if a self-assessment and a test do not match, the test always provides the more correct information.

I was quite closely involved in designing and drafting this feedback, which partly explains my interest in revisiting it in some way even after the DIALANG project ended in 2004, should an opportunity arise. That opportunity materialised in 2006 in the form of a research project called ToLP.

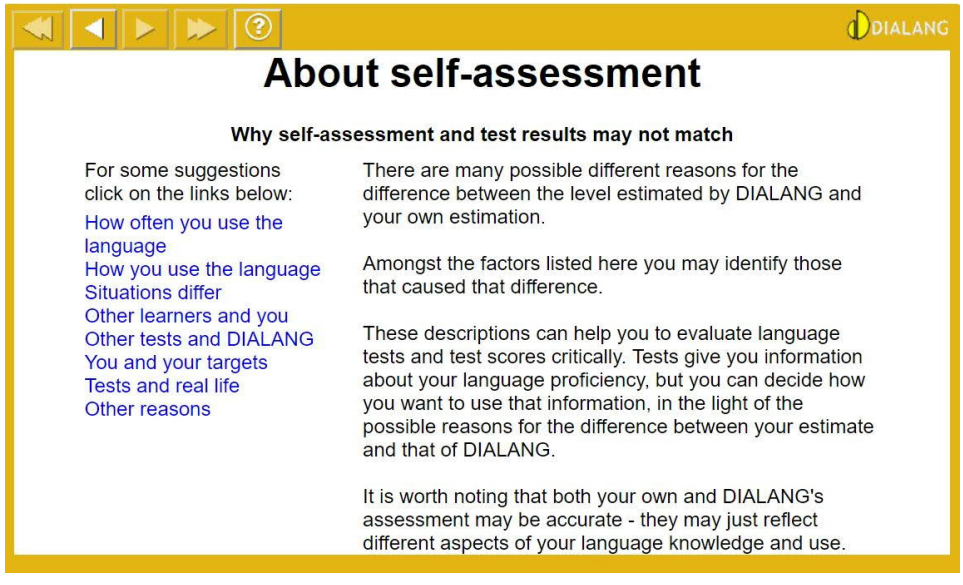


Figure 1. DIALANG feedback explaining possible reasons for a mismatch between self-assessment and test result.

3. The ToLP study

ToLP is an acronym for Towards Future Literacy Pedagogies and refers to an Academy of Finland funded research project at the Centre for Applied Language Studies at the University of Jyväskylä in 2006–2009 (see https://www.jyu.fi/hytk/fi/laitokset/solki/en/research/projects/tolp). The project explored mother tongue and foreign language

literacy practices among Finnish 9th grade students and their teachers in both school and out-of-school contexts. The main part of the study was a large-scale questionnaire-based survey of such practices administered to a statistically representative sample of 9th graders. A similar questionnaire was mailed to their teachers. Here, the focus is on the student survey. The main results of the project are presented in Finnish in Luukka, Pöyhönen, Huhta, Taalas, Tarnanen & Keränen (2008).

3.1 Methods

Participants: The participants were 15-year-old students in grade 9 of the Finnish comprehensive school. A sample of about 2,000 students was selected from the national population of about 55,000 students in grade 9 by using two-staged cluster sampling in which a random sample of schools was first drawn to cover all the regions of the country and different sizes of schools, followed by selecting one intact class from each sampled school. A total of 1,720 students from 101 schools responded to the survey. The response rate was 86 %; the missing 14% consisted of students who were absent from the school or engaged in other activities elsewhere in the school during data collection or who were simply unwilling to fill out the questionnaire.

Questionnaire: The questionnaire covered a wide range of questions about students' reading and writing practices in the school and in their free time, as well as questions about pedagogical practices in the school and students' attitudes towards those practices. The questionnaire was administered in the classroom during one 45 minute lesson; also the following 15 minute break could be used for the purpose if needed (in Finland, lessons last either 45 or 90 minutes and are separated by 15 minute breaks). Data collection was supervised by one of the students' teachers and/or a research assistant working for the project. For more details, see Luukka et al. (2008).

Specific questions about the basis of one's self-assessment: Our previous work on designing DIALANG feedback on self-assessment served as a starting point for designing a sub-set of questions concerning the reasons for the students' view of their proficiency in a foreign language. However, we could not systematically cover all the different factors discussed in *About self-assessment* in DIALANG because the focus of the ToLP project was on somewhat different matters. On the other hand, since the project targeted a specific group of language learners in a particular context, we could include factors that would not have been appropriate on a more general platform such as DIALANG (i.e., questions specific to the school context).

Table 1 reproduces a translated version of the questions of interest for this article. The questionnaire was administered in Finnish, the language of the school and the first language of most of the students.

Table 1. Questions about the basis of students' self-assessment.

*How do you know that you have good or weak skills in a **foreign language**? What affects your view of your language proficiency? **Circle** the most suitable alternative in **each** line.*

		This has affected my view …				
		a lot	**to some extent**	**only a little**	**not at all**	**I do not know**
1.	How easily I learn the language at school.	1	2	3	4	dnk
2.	How I manage to use the language (e.g. abroad, on the Internet, when reading magazines).	1	2	3	4	dnk
3.	What my teachers have said about my language skills.	1	2	3	4	dnk
4.	What my friends and family have said about my skills.	1	2	3	4	dnk
5.	What foreigners I have met have said about my language skills.	1	2	3	4	dnk
6.	How well I succeed in exams at school.	1	2	3	4	dnk
7.	How well I can use the language compared to my classmates.	1	2	3	4	dnk
8.	Something else, what? _____	1	2	3	4	dnk

3.2. Research Questions

The research question of the study reported here were the following:

1) Which factors do the Finnish 9th grade students perceive to affect their view of their foreign language proficiency the most?

2) Are these perceptions related to students' proficiency in their first foreign language?

Students' perceptions were analysed by investigating the distributions of their responses to the questions presented in Table 1 and by conducting multivariate analyses of variance in IBM SPSS (version 24). The measure of the students' foreign language proficiency was their self-reported mark for their first foreign language in their most recent school report on the 7- point scale used in the Finnish comprehensive school. In the scale, the lowest grade 4 indicates a failure to achieve the learning goals for the term and the highest grade 10 denotes excellent achievement (Finnish National Agency for Education, 2004). Because only three students reported the lowest grade (4) as their most recent mark, they were merged with the second lowest grade (5) in the analyses reported below. Since well over 90% of the students had English as their first foreign language, their language proficiency marks refer mostly to that language.

4. Results

The results pertaining to the first research question on the factors that the teenaged learners perceived to have influenced their view of their foreign language skills are presented first followed by the findings that concern the relationship between students' perceptions and their foreign language proficiency. However, before addressing the two research questions, the findings regarding the structure of this battery of questions are reported.

First, correlations between the different questions were computed. It transpired that almost all questions correlated with each other significantly, almost always at $p < .001$ level. The only non-significant correlation occurred between success on the (language) tests or exams in the school and feedback from foreigners. The significant (Spearman rank order) correlations were rather low and ranged between .057 and .429. The fact that even correlations below .1 turned out significant was undoubtedly due to the very large sample size. The strongest correlations were the following:

feedback from friends and family & feedback from foreigners	.429
feedback from friends and family & feedback from teachers	.422
feedback from teachers & examination results-	.419

feedback from foreigners & using language in free time .395

feedback from friends and family & comparison with classmates .317

Overall, the correlational pattern indicates that the questions were mostly tapping rather different aspects of experience. In order to obtain a better picture of the structure of this set of questions, an exploratory factor analysis (Principal Axis Factoring with Promax rotation) was conducted. The analysis suggested that two factors underlie the students' responses, accounting for 54% of variance in their answers. The first factor related to the school and consisted of the questions on teacher feedback, examination results, ease of learning the language in the school, and comparisons with the classmates. The second factor had more to do with language use in free time and it comprised questions about feedback from foreigners, feedback from friends and family, and managing to use the language in free time. The second factor was less clear than the first one: for example, feedback from friends and family loaded on this factor only somewhat more strongly than on the first factor, possibly because many of the students' friends were also their classmates.

To answer the first research question, the distributions, means and standard deviations of the students' responses were investigated. Table 2 describes the distribution of the students' responses to each question. The last two columns in the table display the means and standard deviations (for calculating these, the responses were coded as 1 = *not at all* … 4 = *a lot*; in addition, the *I do not know* responses were excluded because they were considered to be outside the response scale and, therefore, it was not possible to give them any meaningful numerical value). The table shows that the three elements the students thought had most strongly affected their view of their proficiency were how they had managed to use the language in various activities outside the classroom (average 3.48 on the 4-point scale), how they did on the examinations at school (3.27), and how easily they felt they had learned the language at school (3.25).

The factors that the students reported having influenced their views the least included comparison with the other classmates (average 2.63), feedback from foreigners they had met (2.73) and feedback from family and friends (2.74). However, given that 18% of the students had replied "I do not know" and that 14% had chosen "not at all" for the question about feedback from foreigners, it appears that, overall, this type of feedback was the least influential in shaping the students' views of their foreign language proficiency. The high proportion of such answers is likely due to not everybody having had a chance to meet with foreigners in the first place from whom to receive feedback.

Although the distribution of students' answers displayed in Table 2 already gives a fairly clear overall answer to the first research question, the students' mean evaluations were also compared statistically. Given the large sample size (N = 1,343 – 1,621, depending on the variable), it is not surprising that almost all pairwise comparisons of the mean responses to the questions listed in Table 2 turned out to be significant at $p < .001$ level. In fact, the only non-significant pairwise comparisons were

for *ease of learning the language at school* vs *success in the (language) exams at school* and for *feedback from family and friends* vs *feedback from foreigners*.

Table 2. Students' answers to the question asking them to evaluate the degree to which different factors had affected their view of their foreign language proficiency.

	This has affected my view …						
	a lot (4)	**to some extent (3)**	**only a little (2)**	**not at all (1)**	**I do not know**	**Mean (1-4 scale)**	**St. dev.**
1. How easily I learn the language at school.	35%	51%	9%	1%	4%	3.25	.669
2. How I manage to use the language (e.g. abroad, on the Internet, when reading magazines).	57%	32%	8%	1%	2%	3.48	.686
3. What my teachers have said about my language skills.	28%	47%	19%	3%	3%	3.03	.788
4. What my friends and family have said about my skills.	16%	45%	29%	6%	4%	2.74	.813
5. What foreigners I have met have said about my language skills.	23%	28%	18%	14%	18%	2.73	1.05
6. How well I succeed in exams at school.	43%	41%	12%	3%	2%	3.27	.767
7. How well I can use the language compared to my classmates.	18%	36%	29%	11%	6%	2.63	.922

To answer the second research question concerning the relationship between the students' perceptions and their language proficiency, we examined how students with different degrees of proficiency in their first foreign language (as indicated by their most recent mark for that language) differed in their responses to the questions. Table 3 displays the means and standard deviations of the students' replies in each category of proficiency from the lowest (5) to the highest (10).

Table 3 shows a linear increase in the magnitude of values in students' responses for every question: the more proficient students considered these factors to have affected their view of their language proficiency more than the less proficient

students. The last two columns in Table 3 report the correlation (Spearman rank order correlation) between the students' foreign language grades and their responses as well as the difference, for each question, between the least able (mark 5 in the latest school report) and the most able (mark 10) students.

All the correlations in Table 3 were statistically significant at the $p < .000$ level but relatively modest in size. The strongest correlation (.278) between the school mark and the questions was found for the question on how the students had managed to use foreign languages in their free time. This was also the question in which the difference between the responses of the least and most proficient students was considerable, almost one point on the 4-point scale. The question that had the weakest relation to proficiency was the one concerning feedback from friends and family, with the correlation of .174 and a difference of slightly over half a point (0.58 to be precise) on the 4-point scale. All other correlations were in the .19 - .20 region. The question about success in language tests / examinations yielded the largest difference between the most and least proficient learners (0.92 on the 4-point scale) but the correlation between the school mark and responses to this question was only average (.197) compared to the other questions.

Although the students' perceptions of the effects of various factors on their self-assessment increased as their proficiency improved, many of these differences appear rather small and, thus, analyses of their statistical significance are in order. A multivariate analysis of variance (Manova) was conducted to establish whether the differences, overall, across the entire set of questions, were significant. As the analysis indicated that language proficiency (mark in the school report) was significant (Wilks' Lambda = 7.195, $p < .001$, effect size =.038), univariate analyses were carried out to find out to locate the differences.

Table 3. Means and standard deviations of the students' evaluation of the degree to which different factors had affected their view of their foreign language proficiency broken down by students' mark in their first foreign language (mostly English).

Proficiency (mark in the FL)		5	6	7	8	9	10	Difference 5 vs 10	Corr.
1. How easily I learn the language at school.	Mean	2.76	3.08	3.15	3.33	3.33	3.40	0.64	.191
	SD	.751	.669	.674	.602	.690	.613		
2. How I manage to use the language (e.g. abroad, on the Internet,…).	Mean	2.84	3.29	3.37	3.61	3.63	3.74	0.90	.278
	SD	.773	.766	.708	.580	.579	.587		
3. What my teachers have said about my language skills.	Mean	2.69	2.83	2.94	3.00	3.16	3.31	0.62	.209
	SD	.911	.885	.737	.780	.751	.696		
4. What my friends and family have said about my skills.	Mean	2.43	2.63	2.66	2.72	2.91	3.01	0.58	.174
	SD	.957	.807	.754	.825	.761	.761		
5. What foreigners I have met have said about my language skills.	Mean	2.16	2.51	2.67	2.75	3.01	3.01	0.85	.217
	SD	1.048	1.028	.956	1.056	.974	1.030		
6. How well I succeed in exams at school.	Mean	2.57	3.07	3.15	3.30	3.40	3.49	0.92	.197
	SD	.866	.877	.824	.672	.696	.646		
7. How well I can use the language compared to my classmates.	Mean	2.20	2.41	2.53	2.60	2.78	2.93	0.73	.193
	SD	.889	.885	.890	.921	.897	.927		
N		49	160	270	331	355	134		

The last two columns in Table 4 report the results of the univariate analyses (F-values and effect sizes) for each question. The effect sizes (eta squared values) indicate that while language proficiency is significantly associated with the students' perceptions, its effect was rather small. Even the biggest effect size, which was found for managing to use foreign languages in free time, was only .086, which means that only 8.6% of the variance in the students' answers can be attributed to their foreign language proficiency. The second highest effect size was .061 found for how successful the students had been in (language) examinations in the school. For the other factors covered in the questionnaire, the effect sizes were somewhat smaller, ranging from .033 to .046.

Table 4 also displays the results of the pairwise comparisons of the proficiency levels (i.e., different marks). The adjacent levels that were not statistically separable are greyed out. For example, for the first question (ease of learning the language at school), the students' responses at levels (marks) 5 and 6 could not be separated, neither could 6 and 7, which indicates, however, that 5 and 7 could be distinguished. Further, levels 8, 9 and 10 were also indistinguishable, but each of them was distinct from every level below 8. Also levels 7 and 8 could be separated. Overall, then, for this question, those with higher proficiency (marks 8 to 10) formed a statistically distinct group from the less proficient students (from 5 to 7), even if in the latter group a broad distinction could be made between the very weak (5) and the somewhat more proficient (7) learners. For some questions, such as feedback from teachers and comparison with classmates, the picture is somewhat more complex as there are several mutually indistinguishable proficiency groups that partially overlap, but the same principle applies: greyed out levels are not separable but those that lack any colour around them could be distinguished from the grey groups in the current study.

Table 4. Results of the pairwise comparisons of proficiency levels (greyed out levels indicate indistinguishable levels).

Proficiency (mark in the FL)	5	6	7	8	9	10	F	effect size
1. How easily I learn the language at school.							12.568	.046
2. How I manage to use the language (e.g. abroad, …).							24.436	.086
3. What my teachers have said about my language skills.							11.400	.042
4. What my friends and family have said about my skills.							8.925	.033
5. What foreigners I have met have said about my language skills.							11.704	.043
6. How well I succeed in exams at school.							16.847	.061
7. How well I can use the language compared to my classmates.							8.886	.037

5. Discussion

This study was inspired by my involvement in designing the feedback system for DIALANG in the late 1990s and early 2000s. One of the more novel types of its feedback related to self-assessment of language skills. A key aim of DIALANG was, and still is, to support learner agency, autonomy and life-long learning by providing them with feedback on their language skills that also includes advice for further action. Having a chance to try out self-assessment in practice and to learn about it was seen as an important part of this support. While a fair amount of research on self-assessment had been carried out by the time of the DIALANG project (see e.g. the review by Oscarson, 1989, 1997), there were still many unexplored questions and some of the information presented in the feedback was based on expert opinion rather than empirical research. The ToLP project in 2006 – 2009 offered a chance to investigate one aspect of self-assessment in a particular context, namely what factors might underlie language learners' perceptions (i.e., self-evaluations) of themselves as foreign language users, irrespective of whether their perceptions were in accordance with others' views. Thus, a small set of questions targeting possible sources of such perceptions in a school context were designed and administered to a large number of Finnish 9th graders as part of a more comprehensive questionnaire.

The following research questions were of interest in the current study:

1) Which factors do the Finnish 9th grade students perceive to affect their view of their foreign language proficiency the most?

2) Are these perceptions related to students' proficiency in their first foreign language?

As regards the first research question, analyses suggested that the items in the questionnaire formed two broad factors, one that concerned activities that take place in the school (e.g., how well the students do in their studies and on tests, and what feedback they get from the teacher) and another that related to free time language use and feedback from persons other than the teacher. When we examine individual questions, both free time and school related activities were among those that were rated most highly: the highest average was found for managing to use foreign languages in free time followed by two school based activities (how well the FL exams and learning the language in the school had gone). Intuitively, this makes sense because the main foreign language that practically all 9th graders study in Finland is English for which there is a lot of exposure outside the classroom and, therefore, many opportunities to try out one's skills in both on-line and more traditional contexts of language, even at the time of the study in 2006.

Probably one of the most interesting findings concerns the result of the exploratory factor analysis that was performed to investigate the structure of the questionnaire: that separate school and free time related factors could be identified suggests that some students may derive their view of their foreign language proficiency,

and thus, its self-assessment, from what takes place at school whereas for other students their performance in so-called real life activities counts more. A related inference that can be drawn from the results is that the sources of learners' perceptions of their skills are probably quite varied, which is suggested by the rather low intercorrelations of the items in the questionnaire. For many, if not most learners, there may be no one major source of experience that has turned to dominate their perception of themselves as language learners but several factors may in fact play a role in this.

The findings concerning the second research question on the relationship between the students' responses and their foreign language proficiency indicated that the two were significantly correlated. The more proficient the students were, the higher they rated the effect of all the listed types of experience on how they perceived their FL skills. The relationship was not strong, however, and the effect sizes indicated that only 3.3 to 8.6 percent of the variation in the students' responses could be explained by their proficiency in their first foreign language (typically English). The patterns that can be seen in Table 4 suggest that often the most proficient students with marks 8, 9 or 10 in their school report rated the items as more important than those with lower marks. Seen in the context of the research instrument, namely the 4-point Likert scale, it appears that the average ratings by the two most extreme groups of students (mark 4 vs mark 10) could sometimes differ by almost one full scale point.

The reasons why language proficiency was related to the way the students responded to the questions remain unknown, and it is only possible to speculate about the causes. One potential reason relates to the learners' degree of awareness and ability or their willingness to reflect on their skills and learning. If success in language studies, achievement, and self-awareness / self-regulation are associated with each other, then the higher achieving students may also be more aware of themselves as language learners and better able to reflect on which types of experience have had a really significant effect on what they think about their foreign language skills. In contrast, the lower achieving students may have been less used to such metacognitive reflection and, not being quite sure what to answer when requested to do so, chose one of the options indicating a smaller degree of importance.

The limitations of the study include the fact that it was based entirely on a questionnaire survey and lacked a qualitative part which would have shed more light on the students' perceptions. Another limitation concerns the measure of foreign language proficiency used in this study: the students' mark in their first foreign language in their latest school report. First, it was self-reported by the students, which may introduce some inaccuracies. The second, and likely a bigger source of uncertainty, is the lack of standardisation of these teacher-based marks compared with, for example, a standardised language test. One type of uncertainty in this relates to variation between teachers and schools: teachers differ in their grading, as has been shown in national studies of educational achievement in Finland (e.g., Hildén & Rautopuro, 2017). Another type of issue concerns the fact that language marks are not based only on students' language skills but on their achievement in other goals of the curriculum that relate to the target language culture and learning to learn. Furthermore,

the teachers participating in the ToLP study reported clear variation in how much weight they give different factors in at least their final grading at the end of the 9th grade and, thus, presumably also prior to that stage. Some teachers reported even taking the students' diligence, participation, and motivation into account in their grading (Tarnanen & Huhta, 2011).

Further investigation of the topic addressed in this study can be divided in broadly two categories. First, the extensive student questionnaire data collected in the ToLP study could be utilised more comprehensively. For example, to broaden the basis of the evaluation of the students' foreign language proficiency their responses to a range of self-assessment questions could be used in combination of the external measure (mark given by the teacher) that was used in the current study. Another approach to utilising the existing data would be to conduct classification analyses that focus not on the variables (questions) but on the students. Student profiles could be extracted from their responses to the questions investigated in this article in order to find out the number and characteristics of such groups and whether the groups could be related to, for example, students FL proficiency or other relevant data that were collected via the questionnaire.

The second type of research that could build on the current study would be an entirely new investigation that would most likely be more qualitative in nature. Through narrative studies, interviews, learning diaries and other such approaches it could be possible to obtain insights into the kinds of experience that language learners consider important for the formation of their views of what they can do in a foreign language. Such research could also provide insights into how self-assessment works for different types of students in a more general sense, i.e., to what extent it relates to their more general confidence and self-efficacy. It could also increase our understanding of how self-assessment might affect the power relations between language learners and teachers.

To conclude, the current study contributes to the growing body of research on self-assessment of proficiency in a foreign language and hopefully sparks new investigation of the factors that underlie and interact with learners' self-evaluation. It is fair to assume that Sauli Takala would welcome such investigation since self-assessment was one of the many topics that was dear to him not only because of his involvement in the DIALANG project but also because he considered self-assessment to have an important role in language education more generally.

References

Alderson, J.C. (2005). *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. London: Continuum.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching and assessment*. Cambridge: Cambridge University Press.

Gorman, T., Purves, A. & Degenhart, E. (Eds.) (1988) *The IEA Study of Written Composition: The International Writing Tasks and Scoring Scales.* Amsterdam: International Association for the Evaluation of Educational Achievement.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Hildén, R. & Rautopuro, J. (2017). On an equal footing? Comparing school grades and national evaluation results in Finland. In N. Pyyry, L. Tainio, K. Juuti, R. Vasquez, R., & M. Paananen (Eds.). *Changing subjects, changing pedagogies: Diverstities in school and education*, 242–259. Helsinki: Suomen ainedidaktinen tutkimusseura.

Huhta, A. (1987). Kirjoitelmien arviointi. [Assessment of writing] In Takala, S. (Ed.) *Katsauksia kirjoittamiseen.* Julkaisusarja B: Teoriaa ja käytäntöä 5, 25–52. Jyväskylä: Jyväskylän yliopisto: Kasvatustieteiden tutkimuslaitos.

Huhta, A., Luoma, S., Oscarson, M., Sajavaara, K., Takala, S. & Teasdale, A. (2002). DIALANG - A Diagnostic Language Assessment System for Learners. In Alderson, J.C. (Ed*.) Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies*, 130–145. Strasbourg: Council of Europe.

Linnakylä, P., Pollari, P. & Takala, S. (1994). *Portfolio oppimisen arvioinnin tukena* [Portfolio as a way to support assessment of learning]. Jyväskylä: Jyväskylän yliopisto: Kasvatustieteiden tutkimuslaitos.

Little, D. (2005) The Common European Framework and the European Language Portfolio: involving learners and their judgements in the assessment process. *Language Testing, 22*(3), 321–336.

Little, D., Goullier, F. & Hughes, G. (2011). *The European Language Portfolio: The story so far (1991–2011)*. Strasbourg: Council of Europe. Available at https://www.coe.int/en/web/portfolio/history

Little, D. & Erickson, G. (2015). Learner identity, learner agency, and the assessment of language proficiency: Some reflections prompted by the Common European Framework of Reference for Languages. *Annual Review of Applied Linguistics 35*, 120–139

Luukka, M.-R., Pöyhönen, S., Huhta, A., Taalas, P., Tarnanen, M. & Keränen, A. (2008). *Maailma muuttuu - mitä tekee koulu? Äidinkielen ja vieraiden kielten tekstikäytänteet koulussa ja vapaa-ajalla* [The world changes - how does the school respond? Mother tongue and foreign language literacy practices at school and in free time]. Jyväskylä: University of Jyväskylä: Centre for Applied Language Studies.

Finnish National Agency for Education (2004). *National Core Curriculum 2004*. Helsinki: FNAE. Available at https://www.oph.fi/english/curricula_and_qualifications/basic_education/curricula_2004

Oscarson, M. (1989). Self-assessment of language proficiency: rationale and applications. *Language Testing, 6*(1), 1–13.

Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & P. Corson (Eds.) *The Encyclopedia of Language and Education. Volume 7. Language Testing and Assessment*, 175–187. New York: Springer.

Takala, S. (1992). *Virikkeitä uutta kokeilevaan koulutyöhön* [Ideas for innovative trials in education]. Jyväskylä: Jyväskylän yliopisto: Kasvatustieteiden tutkimuslaitos.

Takala. S. (1995). Arviointikulttuurin kehittäminen [Developing assessment culture]. In K. Salmio & K. Lindroos-Himberg (Eds.) *Askelia yleissivistävän koulutuksen arviointiin*, 6–17. Helsinki: Opetushallitus.

Tarnanen, M. & Huhta, A. (2011). Foreign language assessment practices in the comprehensive school in Finland. In D. Tsagari, D. & I. Csepes. (Eds.) *Classroom-Based Language Assessment*, 129–146. Language Testing and Evaluation Series, Grotjahn, R. &. G. Sigott (general eds). Frankfurt: Peter Lang.

Relating Finnish Matriculation Examination grades to the CEFR

Taina Juurakko-Paavola
Tampere University

1. Background

Sauli Takala's work in introducing the CEFR with all its concepts, not just the level descriptions, was very important across all levels of the Finnish school system. Just two examples: he was one of the key people when the target level descriptors and the scale for the Finnish curriculum for comprehensive and general upper secondary education were adapted from the CEFR scale. He was also consulted when the level descriptions for the assessment of the university students' Swedish skills were defined (see Elsinen & Juurakko-Paavola, 2006).

Perhaps one of the most important roles that Sauli Takala had was his involvement in the Finnish Matriculation Examination. Firstly, he was a member of the Matriculation Examination Board for 15 years (1986-2000) and was engaged in a considerable amount of development work during that time. However, even after retiring from the Board he had an important role in developing procedures for relating the Matriculation Examination's language tests to the CEFR. This was undoubtedly one of his many passions: perhaps the main reason for his enthusiasm was that he could see the importance of these comparisons for the international recognition of the results of the language tests of the Matriculation Examination.

It was on Sauli Takala's initiative that the work on linkage was started in 2001 in Finland; he also led the work and conducted the first linkages together with Felianka Kaftandjieva for the English test (Takala & Kaftadjieva, 2002). They applied the same linking procedure to the other languages of the examination in 2004, but these internal reports have not been published. After these first linkage studies, Sauli Takala also wrote the specifications for the language tests in the Matriculation Examination and collected a lot of theoretical background material for the item writers to improve the quality of the tests (Takala, 2006). Updated versions of these documents are still used by the item writers of the examination.

I had the great pleasure of working together with Sauli Takala since 2004 when I became a member of the Examination Board; at that time, he worked as an expert for the language section of the Board. I worked more closely with him since 2012, when we got funding to carry out more linkages and when I worked as a project manager for these activities. In the first project, we related the results of the tests in

English, Swedish, French and German to the CEFR (2012). The findings were published in Finnish (Juurakko-Paavola & Takala 2013) and the results for English were also presented at the EALTA conference in 2014 (Juurakko-Paavola 2014) and are, thus, available at the EALTA website. In the other project, we focused only on English and Swedish (2014), and our main goal was to establish a model which would enable us to estimate how the Finnish Matriculation Examination grades and the CEFR levels relate to each other simply on the basis of the item scores and difficulties even for the examination versions whose items have not been explicitly linked through item centred standard setting procedures. Sauli Takala was very enthusiastic about this project and he made a lot of effort to be able to find a way to do these linkages. We wrote an article reporting this process in Finnish in the autumn of 2015 (published in 2017 as Juurakko-Paavola & Takala, 2017).

2. Main results of two linkage projects

In both linkage projects, the same standard setting method that was developed and successfully used by Kaftandjieva (2010) was applied to set cut scores for the Finnish Matriculation Examination language tests. The method can be described as a cumulative compound method, and it does not require the use of IRT analyses.

Procedures recommended in the Manual (Council of Europe 2009) were applied both times (in 2012 and 2014). About ten experienced panelists (raters and item writers) took part in the standard setting. Evidence collected indicated that internal and procedural validity were good and thus enhanced the validity claim concerning the cut scores (Juurakko-Paavola & Takala, 2013; Juurakko-Paavola & Takala, 2017). Examinee-centred external validation provided further validity evidence in the first project (Juurakko-Paavola & Takala, 2013).

The main research questions in both studies were the same: 1) What is the CEFR level of the tasks in the test? 2) Which CEFR levels do the students achieve? 3) What is the correspondence between the grades given in the Finnish Matriculation Examination and the CEFR levels? The results from both studies indicate that both the English and Swedish tests were estimated to be somewhat easier than the target level set in the Finnish curriculum for the upper secondary school (B2 for English and B1 for Swedish). About 70% of the students reached the target level B2 in English, whereas only about 40% of the students achieved the target level B1 in Swedish.

Seven grades are used in the Finnish Matriculation Exam: *improbatur* (failed), *approbatur*, *lubenter approbatur*, *cum laude approbatur*, *magna cum laude approbatur*, *eximia cum laude approbatur*, and *laudatur*. Until 2014 these grades were given by applying a modified normal distribution. From spring 2014 a new system called the average of standardised total scores has been used (for details see Finnish Matriculation Examination Board, 2018). Neither of these systems is criterion-referenced or competence based.

However, the goal definitions for the Finnish upper secondary school in the national curriculum (Finnish National Agency for Education, 2016) contain target levels adapted from the CEFR levels, as mentioned before. That is why it is very interesting to find out what kind of correspondences there are between the grades in the Matriculation Examination in the English and Swedish tests and the CEFR levels. The results from spring 2014 are illustrated in Table 1. The target levels for the upper secondary school are bolded.

Table 1. Correspondences between the grades in the Finnish Matriculation Examination and the CEFR levels in the English and Swedish tests (spring 2014).

	Appro-batur	Lubenter approbatur	Cum laude approbatur	Magna cum laude approbatur	Eximia cum laude approbatur	Laudatur
English	Strong A2/Low B1	Intermediate B1	**Low B2**	Intermediate B2	Strong B2	Over B2/C1
Swedish	Low A2	Intermediate A2	Strong A2	**Low B1**	Intermediate B1	Strong B1/B2

It can be seen that in the English Examination a student who gets the grade *cum laude approbatur* has reached the target level B2. In the Swedish Examination the students who obtain *magna cum laude approbatur* are at the lower end of the target level B1. This means in practice that in the English Examination most of the students have achieved the target level B2, as mentioned before, whereas in the Swedish test only the three highest grades can be seen as evidence for achieving the target level B1.

After these analyses we concluded that the model presented in Table 2 could work as a starting point for setting cut scores in English and Swedish tests in the Finnish matriculation exam.

Table 2. Cut scores for multiple choice items and writing tasks in English and Swedish for CEFR levels in the Finnish matriculation Examination.

Level		Below A2	A2	B1	B2	C1	C2
Multiple choice (% correct)	English	95–100 %	81–94 %	65–80 %	48–64 %	36–47 %	< 36 %
	Swedish	91–100 %	75–90 %	35–74 %	< 35 %		
Writing (points)	English	< 40 p	42–55 p	58–78 p	80–90 p	92–97 p	99 p
	Swedish	< 51 p	52–62 p	63–86 p	87–99 p		

This model has been used later in another project (Huhta & Juurakko-Paavola 2017), where it was found that, when applied to the 2015 and 2016 tests, this model seems to work very well for the Swedish exam, whereas for the English Examination some modifications would be needed.

3. Conclusions

To relate the results of the Finnish Matriculation Examination to the CEFR levels proved to be very challenging in these projects. The main problem is that all test items are used only once, that is, there is no item bank as in many other high-stakes tests. The reason for this is that the old tests are expected to be released to the schools to be used as learning and testing materials; this happens in all examination subjects and the language tests cannot be an exception. Obviously, an item bank would be needed in the future to ensure the comparability of the tests across different test administrations. An item bank would also considerably reduce the workload needed in the linkage process.

It also became very clear that linking Matriculation Examination grades with the CEFR levels cannot be done automatically, instead a lot of expertise is needed. This was something about which Sauli Takala had a very clear opinion. He emphasised that the experts have to take into account the consequences setting different cut scores can have.

To report test takers' CEFR levels in the Finnish Matriculation Examination certificates the language tests have to meet very high quality standards across all the phases of the test design process. Hopefully, these high standards can be met also in the new digital tests and that, in the near future, the students would also obtain a grade linked to the CEFR levels. This would enhance the transparency of the information reported in the certificates and increase the international usefulness of their grades. That would be something Sauli Takala would have appreciated very much and for which he worked with great enthusiasm for so many years.

References

Council of Europe (2009). *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Strasbourg.

Elsinen, R. & Juurakko-Paavola, T. (2006). *Korkeakouluopiskelijoiden ruotsin kielen taidon arviointi.* Hämeen ammattikorkeakoulu. HAMKin julkaisuja 4/2006.

Finnish Matriculation Examination Board (2018). *Assessment of the examination.* https://www.ylioppilastutkinto.fi/en/assessment-and-certificates/assessment-of-the-examination (Accessed 28.7.2018.)

Finnish National Agency for Education (2016). *National Core Curriculum for General Upper Secondary Schools 2015*. Helsinki: Finnish National Agency for Education.

Huhta, A. & Juurakko-Paavola, T. (2017). *Uusia linkitystuloksia ja ehdotuksia jatkotoimenpiteiksi*. Presentation for the Language Section at the Finnish Matriculation Examination Board 17.3.2017.

Juurakko-Paavola, T. (2014). *The Challenge of Relating National Grading in Examinations to the CEFR.* Presentation at EALTA 2014. http://www.ealta.eu.org/conference/2014/presentations/Taina%20Juurakko-Paavola%20EALTA%202014.pdf (Accessed 28.7.2018)

Juurakko-Paavola, T. & Takala, S. (2013). *Ylioppilastutkinnon kielikokeiden tulosten sijoittaminen Lukion opetussuunnitelman perusteiden taitotasoille.* Ylioppilastutkintolautakunta. http://www.ylioppilastutkinto.fi/images/sivuston_tiedostot/Raportit_ tutkimukset/FI_2013_kielikokeet_taitotasot.pdf (Accessed 28.7.2018.)

Juurakko-Paavola, T. & Takala, S. (2017). Kohti kriteeriperustaista arviointia ylioppilastutkinnon kielikokeissa. In V. Britschgi & J. Rautopuro (eds). *Kriteerit puntarissa*. Kasvatusalan tutkimuksia 74. Jyväskylä: Suomen kasvatustieteellinen seura, 41–62.

Kaftandjieva, F. (2010). Methods for setting cut scores in criterion-referenced achievement tests. A comparative analysis of six recent methods with an application to tests of reading in EFL. EALTA. http://www.ealta.eu.org/resources.htm (Accessed 28.7.2018.)

Kaftandjieva, F. & Takala, S. (2002). *Relating the Finnish Matriculation Examination English Test Results to the CEF Scales.* Paper presented at the Seminar on Linking Language Examinations to CEFR. Helsinki, June 31−July 2, 2002.

Takala, S. (2006). *Ylioppilastutkinnon kielikokeiden laadinnan opas.* Ylioppilastutkintolautakunta (internal document).

Nuoresta opiskelijasta alansa asiantuntijaksi

Erkki Kangasniemi
Koulutuksen tutkimuslaitos
(Finnish Institute for Educational Research)
University of Jyväskylä

Abstract

In his article "From a young student to an expert in his field", Erkki Kangasniemi, former director of the Finnish Institute for Educational Research (FIER), gives an account of Sauli Takala as a student and a colleague. Sauli, Erkki and a close friend and colleague of Sauli's, Liisa Havola, all started their studies in 1961 at the University where Sauli studied English and Swedish. Sauli was a very diligent student who wanted to achieve the highest grade on every examination he took – and he was usually successful in this. On one occasion, Sauli failed to get the grade he thought he deserved on a phonetics examination; he took the study books with him to the lecturer and by showing how he had used them to answer the questions he succeeded in persuading the lecturer to raise his grade to a level that satisfied Sauli.

Sauli Takala graduated from the university in 1970. To fund his living expenses during his studies, he worked as a part-time language teacher in different schools on several occasions. He also completed pedagogical studies to have the formal language teacher's qualifications. Sauli started his career as a researcher already while studying (which probably explains why it took him almost ten years to graduate). From 1966 onwards Sauli worked at the FIER in various projects that paved the way for the introduction of comprehensive education in Finland (that happened in stages in the 1970s). Sauli specialised in investigating language education and curriculum design. In those days, as Erkki Kangasniemi reminisces, the national educational authorities did not consider researchers and research findings as something very useful for educational planning. Research results were often ignored, if they were at odds with what the existing laws and regulations about education stated.

Starting in 1971, both Sauli and Erkki began to work as research assistants in the new unit within the FIER that focused on conducting evaluations of educational achievement in the schools that trialled comprehensive education, often comparing the findings with those obtained in the traditional 'middle schools' that represented the old, selective educational system. The introduction of the comprehensive school in Finland was a highly political issue: there were widespread arguments that not everybody can study according to the same curriculum till grade 9 (which marks the end of compulsory education in the new system) and that not everybody can learn a foreign language

(which became a compulsory subject in the comprehensive school). Therefore, it was important to have empirical data to counter the claims that the introduction of comprehensive education would drastically lower the standards and achievement. Apparently, at least some politicians and educational authorities had started to pay attention to research by that time!

Sauli continued to work on the educational evaluation of foreign and second languages in the 1970s and 1980s at the FIER but he was also appointed head of the publishing unit of the Institute. Since the 1970s, Sauli Takala was engaged in many national level committees and working groups that focused on language questions such as analysing the needs for different foreign languages at the national level.

In the early 1980s, the FIER became involved in the planning and conducting of the international comparative study of writing by the IEA (International Association for Evaluation of Educational Achievement) and Sauli became the international coordinator of the study with Alan Purves and moved to the University of Illinois at Urbana-Champaign in the USA. While coordinating the study, Sauli also completed his doctoral studies and graduated in 1984.

Sauli returned to Jyväskylä and the FIER at the end of 1984 and began to publish actively on language learning, teaching and assessment. He was appointed to the Matriculation Examination board (the ME is the final school leaving examination for the general/ academic upper secondary education) to represent the English language. Since the 1980s, Sauli supervised and examined numerous doctoral dissertations both in Finland and abroad, particularly in the other Nordic countries. He also held a temporary Professorship in Applied Linguistics before moving to the Centre for Applied Language Studies at the University of Jyväskylä in the mid-1990s, first as a Senior Researcher and then as a Research Professor.

1. Sauli Takala opiskelijana

Edellisen vuoden ylioppilaista olivat useimmat pojat syksyllä 1961 suorittaneet asevelvollisuutensa ja olivat valmiina aloittamaan opiskelun yliopistossa tai korkeakoulussa. Niinpä syksyllä 1961 Isonkyrön yhteiskoulun vähäkyröläinen ylioppilas Sauli Jaakko Takala monien muiden, muun muassa Liisa Havolan ja Erkki Kangasniemen, tavoin aloitti opintonsa Kasvatusopillisessa korkeakoulussa Jyväskylässä. Sauli ja Liisa aloittivat englannin kielen opintonsa; Sauli opiskeli myös ruotsin kieltä. Erkki aloitti valmistautumisensa opettajan tehtävään.

Me siis aloitimme opinnot Jyväskylässä yhtä aikaa. Liisa tutustui Sauliin heti syksystä 1961 alkaen. Ei liene epäselvää oliko aloitteen tekijänä Sauli vai Liisa. Englannin kielen uusia opiskelijoita oli nelisenkymmentä, joista neljä oli miehiä. Liisa esittäytyi Saulille ja siitä ystävyys alkoi. Liisa asui siskonsa kanssa yksiössä Puistokadulla, Sauli alivuokralaisena omakotitalon yläkerrassa. Päivät kuluivat opiskellen. Sauli oli ahkera ja menestyvä opiskelija, joka ei ehtinyt osallistua ylioppilaskunnan toimintaan. Hän tarjoutui Liisalle ja hänen siskolleen keskustelukumppaniksi opiskeluun liittyvissä

asioissa. Yleensä keskiviikkoisin Sauli kävi tyttöjen asunnolla ja preppasi heitä opinnoissa. Sauli kertasi tulevan tentin asioita tyttöjen kanssa, jolloin he oppivat asiat, sillä kielenopetus yliopistossa oli enemmän tai vähemmän lapsen kengissä. Näin Sauli paikkasi opetuksessa jääneitä aukkoja. Vastineeksi tytöt tarjosivat keskiviikkoisin Saulille erilaisia aterioita asunnollaan.

Sauli sai opintojensa kuluessa parhaan arvosanan kaikista tenteistä. Lehtori Leho Võrk piti kaikille kielen opiskelijoille yleisen fonetiikan kurssin. Leho Võrk oli ankara ja hiukan omintakeinen opettaja. Opiskelijat suorastaan pelkäsivät hänen tenttiään ja joidenkin opinnot keskeytyivät siihen, että lehtori ei hyväksynyt tenttiä. Sauli sai fonetiikan tentistä mielestään niin huonon arvosanan, että hän meni lehtori Võrkin vastaanotolle tenttikirjat mukanaan ja kirjoista osoitti, miten hän oli vastannut tenttikysymyksiin. Neuvottelun tuloksena Saulin arvosana fonetiikan tentistä nousi häntä tyydyttävästi.

Kuten sanottu, Sauli menestyi opinnoissaan hyvin. Vain kerran opintojensa historiassa hän sai huonomman arvosanan tentistä kuin Liisa Havola; kyseessä oli Amerikan historian tentti. Liisa oli ollut kouluaikana stipendiaattina yhden vuoden Amerikassa ja opiskellut siellä Amerikan historiaa, joten hänellä oli tavallaan etulyöntiasema.

Sauli otti filosofian kandidaatin eli maisterin paperit Jyväskylän yliopistosta vuonna 1970. Jyväskylän Kasvatusopillinen korkeakoulu oli muuttunut Jyväskylän yliopistoksi vuonna 1966, joten hajurakoa korkeakouluun tuli hänen tutkintonsa osalta nelisen vuotta.

Me monet opiskelimme opintolainan turvin. Toki kesällä mahdollisuuksien mukaan olimme töissä ja säästimme palkasta tulevaa lukuvuotta varten. Sauli tienasi opiskelurahoja tutustumalla mahdolliseen tulevaan ammattiinsa. Hän oli vuonna 1964 yhteensä neljä kuukautta tuntiopettajana Isonkyrön yhteiskoulussa ja nuorempana lehtorina Vähänkyrön yhteiskoulussa sekä vuonna 1965 pari kuukautta lehtorina Jyväskylän kunnallisessa keskikoulussa. Lisäksi hän auskultoi 1970-luvulla lukukauden verran saadakseen erivapauden oppikoulujen vanhemman ja nuoremman lehtorin virkoihin. Sauli piti näin huolta siitä, että yliopistoon voidaan rekrytoida työkokemusta omaavia lahjakkaita henkilöitä.

2. Sauli Takala tutkijana

Se meille on epäselvää, miten kieliä opiskellut mies ajautui Kasvatustieteiden tutkimuslaitokseen (KTL) tutkijaksi. Muistaaksemme syksystä 1966 alkaen Sauli työskenteli ns. kouluhallituksen tutkijaryhmässä Pentti Pihasen ja Pekka Käpin kanssa. Kouluhallituksen rahoittamana he tekivät tutkimus- ja selvitystyötä koulu-uudistuksen toimeenpanoa varten. Sauli keskittyi kielenopetuksen kehittämiseen; opetettavista kielistä ei ollut vielä päätetty ja niinpä Sauli englannin ja ruotsin lisäksi avusti saksan kielen lehtoria saksan oppimateriaalin laatimisessa. Tuohon aikaan tutkijat olivat outoa porukkaa kouluhallituksen virkamiehille ja tutkimustulokset vielä oudompaa kuultavaa. Tutkijoiden tuloksia tyrmättiin siltä pohjalta, että laki ja asetus sanoo asiasta niin ja

niin. Koulun tulee toimia niin kuin laki sanoo; kaikki muu oli virkamiesten mielestä lainvastaista. Itse työskentelin tuohon aikaan opettajanvalmistuslaitoksessa ja joskus ihmettelin itsekin, että mitä porukkaa nuo kolme muuten niin mukavaa miestä ovat. Kokeiluperuskouluja varten valmistettiin v. 1968 väliaikainen opetussuunnitelma ja luulen, että ylitarkastaja Jouko A. Räihä tuohon aikaan veti Saulin mukaan kielenopetuksen uudistamishankkeisiin.

Vuoden 1971 alusta käynnistettiin Kasvatustieteiden tutkimuslaitoksessa koulututkimusosasto, joka keskittyi koulu-uudistusta koskevaan tutkimukseen. Kouluhallitus halusi rahoittaa tutkimusta peruskoulu-uudistuksen tueksi. Tällöin Sauli ja minä (Erkki) olimme molemmat tutkimusassistenttina koulututkimusosaston eriyttävän opetuksen tutkimusryhmässä. Siinä ryhmässä Sauli, Puron Jussi ja Koppisen Leena suorittivat vieraan kielen, matematiikan ja äidinkielen oppimistulosten arviointia kokeiluperuskouluissa ja Piipon Teuvo ja minä pyrimme selvittämään eriyttämistä ilmiönä. Arviointitulokset olivat kouluhallitukselle erittäin tärkeitä, koska peruskoulua vastustettiin muun muassa siksi, että se heikentää oppimistuloksia. Kokeiluperuskouluista saadut tulokset eivät vahvistaneet tätä pelkoa. Tuolloin kokeiluperuskoulujen tuloksia saatettiin verrata myös rinnakkaiskoulun eli keskikoulun ja kansalaiskoulun vastaavien luokka-asteiden oppilaiden tuloksiin. Oppimistulosten arviointi oli meidän tutkimusryhmämme jäsenille tuttua, koska laitos oli aikaisemmin jo osallistunut IEA:n (International Association for Evaluation of Educational Achievement) kansainväliseen arviointitoimintaan. Sopivasti ryhmämme toiminnan alkuaikoina v. 1971 oli Ruotsissa IEA:n puitteissa järjestetty ”Grännan seminaari”, johon Saulikin osallistui.

Oppimistulosten arviointi jatkui vuodesta toiseen 1970-luvun alkuvuosina. Kun arviointi tuli tärkeäksi osaksi koulu-uudistusta, niin perustettiin Kasvatustieteiden tutkimuslaitoksen koulututkimusosastoon ns. arviointiyksikkö eli koulukoetoimisto. Sikäli kun muistan, eriyttävän opetuksen tutkimusryhmä menetti Saulin tutkimuslaitoksen tietopalveluosaston johtoon. Olimme erittäin tyytymättömiä, kun tutkimusryhmä menetti Saulin. Kuitenkin tulimme Saulin lähdön jälkeenkin hyvin toimeen ja Liisa Havola liittyi pian vastaperustettuun koulukoetoimistoon. Ollessaan tietopalveluosaston johtajana Sauli vielä oli yhteydessä vieraan kielen arviointitoimintaan. Hänen päätehtävänsä oli kuitenkin kehittää laitoksen julkaisutoimintaa ja tiedottamista. Hän muun muassa luki kaikki laitoksen julkaisut ennen niiden painamista. Kansallinen ja kansanvälinen tiedottaminen oli monimuotoista ja kansainvälisessä tiedottamisessa Saulin kielitaito oli ehdoton etu laitokselle.

Todettakoon tässä vaiheessa jotakin Saulin tutkijan asiantuntemuksen hyödyntämisestä. Sauli oli monen kouluhallituksen työryhmän jäsen tuoden tutkimuksen näkökulmaa työryhmien työskentelyyn. Hänellä oli toiminnalliset suhteet kielen ylitarkastajiin, Jouko A. Räihään ja Rauha Petroon. Kielenopetuksen perustavoitteiden määrittelyssä Sauli antoi tutkimuksen ja oman oppineisuutensa näkyä. Perustavoitteet olivat joidenkin mielestä sellaisia, mitä pitää kaikille opettaa, toisille taas ne olivat sellaisia, joita kaikkien tulisi oppia. Kielenopetus ja oppiminen painottui myös kansliapäällikkö Jaakko Nummisen johtaman kielitarvetoimikunnan työssä. Sen sihteerinä Sauli teki huomattavan työn selvittäen mitä kommunikaatiovälineitä globaalissa maailmassa

eri aloilla tarvitaan ja joihin tarpeisiin koulun kieliohjelman tulisi vastata. Sauli toimi myös opettajana ja luennoitsijana monissa seminaareissa ja koulutustilaisuuksissa.

Kun laitos muutti 1970-luvun lopulla uusiin tiloihin, niin Sauli oli edelleen tietopalveluyksikön johdossa, mutta veri veti voimakkaasti myös englannin kielen oppimistulosten arvioimiseen. Kun tutkimuslaitoksessa valmistelimme IEA:n yleiskokousta (General Assembly) pidettäväksi Jyväskylässä, niin kokouksen ohjelmaan tuli myös uuden hankkeen suunnittelu. Yleiskokous järjestettiin elokuussa 1980 ja siinä päätettiin IEA:n kansainvälisen kirjoitelmatutkimuksen suunnittelun aloittamisesta. KTL tuli yhdeksi tutkimuksen vastuuyksiköksi. Lisäksi henkilökysymyksistä keskusteltaessa päädyttiin siihen, että Sauli Takalasta tuli kirjoitelmatutkimuksen kansainvälinen koordinaattori yhdessä Alan Purvesin kanssa. Vuoden 1981 helmikuusta alkaen Sauli siirtyi IEA:n palvelukseen aloittaen työnsä KTL:ssa ja siirtyi sitten USA:han Illinois'n osavaltioon. Hänen työpaikkansa oli University of Illinois'n eräällä laitoksella. Tällöin Saulille avautui mahdollisuus katsoa pitkälle eteenpäin eikä pelkästään jalkoihinsa. Kirjoittamisen käsitteellistäminen ja teoreettinen auki kirjoittaminen tuotti ajatuksia ja mielikuvia, joita muut eivät vielä olleet esittäneet. Kansainvälisen koordinaattorin työskentely yhdessä Illinois'n yliopiston joidenkin henkilöiden kanssa oli erittäin merkittävää tutkimuksen suorittamisen kannalta. Siltä pohjalta Sauli saattoi kirjoittaa papereita kirjoitelmatutkimuksen kansainväliselle ohjausryhmälle ja tehostaa ohjausryhmän työskentelyä. Uusiseelantilainen Robert Garden, joka oli mukana IEA:n toisessa kansainvälisessä matematiikkatutkimuksessa, sattui muutaman kerran olemaan läsnä kirjoitelmatutkimuksen ohjausryhmän kokouksissa ja hän kehui Saulin ja kirjoitelmatutkimuksen kansainvälisen ohjausryhmän työskentelyä tehokkaaksi.

Illinoisissa työskennellessään Sauli myös suoritti PhD –tutkinnon (Doctor of Philosophy) vuonna 1984 ennen kuin hän palasi Suomeen. Muistelen, että vuosi oli 1984 ja että itse kävin Illinoisissa vuonna 1982 IEA:n toisen kansainvälisen matematiikkatutkimuksen työseminaarissa ja olin majoittuneena Saulin luona.

Vuoden 1984 lopulla Sauli oli jälleen Suomessa ja Kasvatustieteiden tutkimuslaitoksessa. Tällöin hän työskenteli kansainvälisen kirjoitelmatutkimuksen puitteissa kirjoitellen ja viimeistellen artikkeleita julkaistavaksi. En muista oliko hän mukana kouluttamassa kirjoitelmatutkimukseen osallistuvia opettajia oppilaidensa kirjoitelmien arvioimiseen. Saulin johdolla oli kansainvälisessä ohjausryhmässä määritelty selkeät arviointikriteerit, joita kussakin maassa sovellettiin. Sauli tavallaan toimi kansainvälisen kirjoitelmatutkimuksen "säkkinä", ideoijana ja muut "säkin suuna" levittäen Saulin ajatuksia kuulijakunnallensa.

Kansainvälinen kirjoitelmatutkimus laajensi Saulin intressiä kielen oppimiseen, opetukseen ja kielitaidon arviointiin. Sen jälkeen hän oli asiantuntijana monissa eri tehtävissä. Hän oli ylioppilastutkintolautakunnan englannin kielen jaoksen jäsenenä ja pyrki tehokkaasti kehittämään ylioppilastutkinnon kielitaidon arviointia; hän oli useiden väitöskirjojen ohjaajana, tarkastajana ja vastaväittäjänä niin kotimaassa kuin ulkomaillakin. Sauli oli myös alansa kansainvälisten järjestöjen asiantuntijana ja joidenkin järjestöjen puheenjohtajana.

Saulilla oli Jyväskylän yliopistossa vetoa eri laitoksille. Hän hoiti muun muassa 1980-luvun lopulla soveltavan kielitieteen ja puheentutkimuksen professuuria jonkin aikaa. Vuonna 1989 huhu kierteli KTL:ssa, että Sauli on lähdössä laitoksesta ja vuoden 1996 alussa hän siirtyi pitkän harkinnan jälkeen Soveltavan kielentutkimuksen keskukseen, ensin erikoistutkijaksi ja sitten parin vuoden jälkeen tutkimusprofessoriksi. Näin valmis, pitkälle kouluttautunut mies vietiin pois KTL:stä paremmille paikoille, paremmille palkoille

Monet erilaiset tehtävät Kasvatustieteiden tutkimuslaitoksessa (1966-1989) ja erityisesti IEA:n kansainvälisen kirjoitelmatutkimuksen kansainvälisen koordinaattorin tehtävät 1980-luvulla antoivat Sauli Takalalle vankan perustan, jonka varassa hänestä kehittyi kielididaktiikan eli vieraankielen opettamisen, opiskelemisen, oppimisen ja arvioinnin todellinen asiantuntija. Saulin julkaisuista ja elämäntyöstä saa hyvän kuvan hänen ylläpitämiltään verkkosivuilta osoitteessa https://kiesplang.fi/.

Kielikylpykahvila ruotsinopiskelijoiden informaalina oppimisympäristönä

Carola Karlsson-Fält
Turun yliopisto
(University of Turku)

Abstract

A language immersion café as an informal learning environment for students in the Swedish language

The qualification requirements of Finnish-speaking students in Finland include the requirement of Swedish language proficiency, which students demonstrate in both an oral and a written exam. The required CEFR (Common European Framework of Reference) level has remained the same, at B1 or B2, although since 2005 Swedish is no longer an obligatory subject in the national Matriculation Examination at the end of general upper-secondary education. The scope of the Swedish courses at university has also remained unchanged, although the starting level of most students at the beginning of the course is considerably weaker than earlier. The starting level of the students, the diversity of their disciplines studied and the high number of students pose challenges for Swedish language courses.

To improve oral language skills, in particular, it has become necessary to find support for learning from various informal learning environments. This article focuses on how students who participated in a discussion group at a language immersion café experienced the use of Swedish outside the classroom. The discussion group meets once a week in the Finnish Swedish information and culture centre *Luckan*. Participation in the discussion group has been integrated into the course.

The data, which were gathered over two and a half years (Spring 2016 – Spring 2018), consist of quite short informal student reports written about the visits to the discussion group. The reports were written in Swedish by students from the Faculty of Humanities and the Faculty of Social Sciences of the University of Turku. In total there were 190 reports. In the analyses, the expressions used in the reports were summarised and grouped according to use of similar expressions. After the analysis, four different classes of expressions were distinguishable with the class expressing emotions being the most abundant. The expressions in the other three classes concentrated on reflection of language usage and language learning, the limitations of the classroom, and the compulsory participation in the discussion group. The data reveal that almost all students experienced nervousness and stress as they had to participate in a group discussion, but many also experienced relief and joy as they noticed that they could get by with their language skills. A learning environment without a teacher and supervision

was experienced as motivating. The authenticity of the learning situation helped the students to reflect their language use and to discover an opportunity for comprehensive language use. The compulsion to participate in the discussion group was experienced as positive.

1. Johdanto

Kieltä on aina opiskeltu luokkahuoneen ulkopuolella tapahtuvaa käyttöä varten, käytettäväksi elävässä elämässä, mutta sitä on pitkään opiskeltu pelkästään luokkahuoneessa, joka oppimisympäristönä on tarkoitukseen hyvin rajallinen. Yliopistotutkintoon kuuluvalla ruotsin kielen kurssilla rajoittavina tekijöinä on tavallisesti koettu samassa ryhmässä opiskelevien opiskelijoiden alojen moninaisuus ja opiskelijoiden suuri määrä (ks. Karlsson-Fält – Maijala 2007, s. 333; Kuokkanen-Kekki & Niedling 2011, s. 42; Richards 2015, s. 5–6), mutta ruotsin kielen taitojen jatkuvasti heikennyttyä kurssin rajoittaviksi tekijöiksi ovat tulleet myös samalla kurssilla opiskelevien opiskelijoiden eri taitotasot. Rajallisuus voi edellä ilmenneiden seikkojen lisäksi kuitenkin ilmetä myös oppimistilanteessa siten, että luokkahuoneen ainoa sujuvasti kohdekieltä puhuva on opettaja ja kielen harjoittelu pelkistyy tekstin tai tietyn aihepiirin käsittelyyn sen sijaan, että kielellä voisi toimia autenttisessa tilanteessa (vrt. Krashen 1981, s. 137).

Tänä päivänä luokkahuoneessa tapahtuvaa säänneltyä kielenoppimista, ns. formaalia oppimista tukee ehkä entistä enemmän ja monimuotoisemmin luokkahuoneen ulkopuolella tapahtuva ns. informaali oppiminen. Media, tietoverkot ja erilaiset sähköiset oppimisalustat mahdollistavat kielen oppimisen myös luokkahuoneen ulkopuolella. Sähköiset alustat antavat joustavan ja monipuolisen mahdollisuuden oppia kieltä henkilökohtaisten tarpeiden mukaan ajasta ja paikasta riippumatta (ks. Richards 2015, s. 5–6). Suullisen kielitaidon ja keskustelutaidon kehittämiseksi näiden alustojen rinnalle on kuitenkin yhä enemmän alettu kaivata mahdollisuuksia kohdata kasvokkain kohdekieltä puhuvia autenttisissa kielenkäyttötilanteissa.

Perinteisesti opiskelijoiden puhumis- ja keskustelutaitoa harjoitellaan luokkahuoneessa pari- ja ryhmäkeskusteluissa. Suullista kielitaitoa kohennetaan myös erilaisilla tandem- ja kielikaveriharjoituksilla, joissa osapuolet ovat natiiveja kielenkäyttäjiä ja oppivat toisiltaan kielenkäyttöä (Benson 2013, s. 131; Krashen 1981, s. 105–106; ks. Kuokkanen-Kekki & Niedling 2011, s. 39). Kohdekieltä puhuvien vierailut luokkahuoneessa sekä osallistuminen vieraskielisille luennoille ovat myös olleet suosittuja tapoja oppia kieltä. Näissä oppimistilanteissa opettajan tehtävänä on yleensä organisointi, kontrollointi ja arviointi. Turussa Kielten ja viestinnän keskuksen ruotsinopiskelijat ovat osallistuneet myös ruotsinkielisille opastetuille käynneille Turun linnaan ja tuomiokirkkoon sekä käyneet ruotsinkielisissä jumalanpalveluksissa ja elokuvissa. Nämä oppimistilanteet vahvistavat jossakin määrin opiskelijan taitoa ymmärtää kieltä, mutta eivät aina anna tilaisuutta varsinaiseen puheharjoitteluun.

Luokkahuoneen ulkopuoleiset, mahdollisimman autenttiset oppimistilanteet, joissa opiskelija voi puhua vierasta kieltä ilman opettajan tukea, tehtäviä ja arviointia,

ovat opiskelijan kielenkehitykselle tärkeitä. Näissä opiskelija saa rohkeutta hyödyntää erilaisia oppimistilaisuuksia ja hän oivaltaa samalla elinikäisen oppimisen tärkeyden. Opettajan taas on tärkeää saada tietoa siitä, miten opiskelijat nämä tilanteet kokevat, jotta opetusta voidaan kehittää kannustamaan tällaiseen oppimiseen jatkuvasti. (Vrt. Karlsson-Fält & Maijala 2007, s. 332.)

Tässä artikkelissa tarkastelen Turun yliopiston Kielten ja viestinnän keskuksessa ruotsin kieltä opiskelevien opiskelijoiden kokemuksia keskusteluharjoituksista, joihin he osallistuivat Turussa toimivassa kielikylpykahvilassa. Aluksi, luvussa 2 selvitän formaalin ja informaalin oppimisen käsitteitä sekä näiden liittymistä elinikäisen oppimisen ideologiaan. Luvussa 3 kerron lyhyesti opiskelijoiden ruotsin kielen kurssista yliopistossa ja kielikahvila Luckanista. Luvussa 4 paneudun aineistoon ja sen analyysiin ja luvussa 5 vertailen lyhyesti kielikylpykahvilaa ja muita opiskelijoiden hyödyntämiä informaaleja oppimisympäristöjä. Luvussa 6 pohdin tutkimuksen antia opiskelijoille ja opettajalle.

2. Formaali ja informaali oppiminen

Yksinkertaisin tapa määritellä formaali ja informaali oppiminen on tehdä se sen mukaan opitaanko jotakin intentionaalisesti muodollisissa oppilaitoksissa kuten kouluissa, yliopistoissa ja muissa oppimiskeskuksissa, vai opitaanko jotakin sekä intentionaalisesti että toisinaan myös ei-intentionaalisesti tällaisten laitosten ulkopuolella. Formaalille oppimiselle on olennaista valmis opetussuunnitelma, jota koulutettu opettaja noudattaa, sekä opiskelijat, joita arvioidaan. Opiskelijat tietävät mitä heidän oletetaan oppivan ja he hyväksyvät sen, että oppilaitos tietyssä määrin kontrolloi heitä. Informaalia oppimista on sitä vastoin kaikki muu oppiminen. (Hager & Halliday 2009, s. 1–2; s. 29–31; Richards 2015, s. 5–6; ks. myös Krashen 1981, s. 40; Singh 2015, s. 45.) Formaalin ja informaalin oppimisen lisäksi voidaan puhua non-formaalista oppimisesta, jolla tarkoitetaan formaalin oppimisen täydentämistä tai sille annettavia vaihtoehtoja. Non-formaali oppiminen voi olla säänneltyä kuten oppilaitoksissa, mutta luonteeltaan joustavampaa. Toisinaan non-formaalin ja informaalin oppimisen käsitteitä käytetään keskenään vaihtoehtoisesti. Kaikki nämä oppimisen muodot tulisi nähdä kokonaisuutena ja jatkumona eikä erillisinä osa-alueina, ja taidot, joita ei ole saavutettu formaalissa oppimisessa, tulisi tunnistaa. (Hager & Halliday 2009, s. 27; Singh 2015, s. 19–20; Siurala 2002, s. 71–72; vrt. Dewey 1948, s. 90–91.) Tässä artikkelissa käytän termejä formaali ja informaali oppiminen; informaalia oppimista luokkahuoneen ulkopuolisena oppimisena.

Kiinnostus informaaliin oppimiseen on kasvanut elinikäiseen oppimiseen kohdistuvan kiinnostuksen myötä. Ajatukseen elinikäisestä oppimisesta sisältyvät erilaiset oppimisympäristöt ja oppijan sosiaaliset, taloudelliset sekä henkilökohtaiseen kehitykseen liittyvät tavoitteet. Elinikäinen oppiminen ymmärretään koko elämän pituisena jatkuvana oppimisena (life-long learning) ja siihen liittyy myös oppimisen yhdis-

täminen ihmisen koko muuhun elämään - oppimiseen kotona, kouluissa, työssä, yhteisöissä, arkielämässä, vapaa-aikana jne. (life-wide learning). (Singh, 2015, s. 19.) Elinikäisen oppimisen kokonaisuuteen katsotaan nykyään kuuluvaksi myös arvojen oppiminen: uskonnolliset, moraaliset, eettiset ja sosiaaliset arvot ohjaavat oppijoita siinä, mihin he uskovat ja miten he toimivat itseään ja muita kohtaan. Tässä oppimisessa olennaista on kieli. Ihmisen tulee ymmärtää kuinka käyttää kieltä eri rooleissa kuten vanhempana, mentorina, työntekijänä jne. Jokainen rooli vaatii monenlaista puhetta tai kielellistä esittämistä (life-deep learning). (Banks, Ball, Gordon, Gjutierrez, Heath, Lee, C., Lee, Y., Mahiri, Nasir, Valdes & Zhou 2007, s. 12; ks myös Huhta 1993, s. 90–91.)

Elinikäisen oppimisen edellytyksenä pidetään opiskelijan kykyä itseohjattuun opiskeluun. Oppiminen voidaan nähdä oppijan inhimillisenä, kokemukseen pohjautuvana kasvuna. (Siekkinen 2017, s. 46, 73). Oppimistilanteessa opiskelijan tulee saada olla kokonaisena ihmisenä: ei ainoastaan kognitiivisena toimijana, vaan myös emotionaalisena ja sosiaalisena ihmisenä. Tunteiden mukanaolon on todettu johtavan tehokkaampaan oppimiseen. (Kaikkonen 2000, s. 55). Kielenopetuksessa korostuvat nykyään oppijakeskeisyys ja -lähtöisyys, autonominen oppiminen, oppimaan oppiminen, kontekstuaalisuus sekä autenttisuus: aidot, mielekkäät kielenkäyttötilanteet ja kielen funktionaalisuus. Nykyisen sosiokulttuurisen oppimiskäsityksen mukaisen opetuksen tulisi tukea tilanteista oppimista, joka perustuu oppijoiden vuorovaikutukseen ainutkertaisessa tilanteessa. Oppimistilanteen huomioiminen oppimisen lähtökohtana on olennaista, ei menetelmän. (Järvinen 2014, s. 111–112; vrt. Takala, 1992, s. 14.) Hyvässä oppimistilanteessa opiskelija pääsee monipuolisesti kosketuksiin kielen kanssa, omaksuu kieltä palkitsevassa vuorovaikutuksessa muiden kanssa kehittyen samalla itse kokonaisvaltaisesti (van Lier 2000, s. 254–255).

3. Oppiminen yliopiston kielikurssilla ja kielikylpykahvilassa

Moni opiskelija, joka suorittaa yliopistossa tutkintoonsa kuuluvaa ruotsin kielen kurssia kertoo, ettei ole koskaan saanut tilaisuutta käyttää ruotsia luokkahuoneen ulkopuolella tai sen käyttö on jäänyt hyvin vähäiseksi. Myös puhutun kielen käyttö luokkahuoneissa on opiskelijoiden mielestä usein ollut vähäistä. Tutkintoon kuuluva ruotsin kielen kurssi yliopistossa on integroitu kurssi, jolla harjoitellaan sekä puhuttua että kirjoitettua kieltä. Se asettuu eurooppalaisen viitekehyksen taitotasolle B1-B2. Taitotasolla B1 puheen tulisi yleisesti ottaen olla *melko sujuvaa*, taitotasolla B2 *sujuvaa*. Taitotasolla B1 opiskelijan tulee esimerkiksi valmistautumatta pystyä osallistumaan keskusteluun aiheista, jotka liittyvät arkielämään ja ajankohtaisiin asioihin, ja taitotasolla B2 hänen tulee osata viestiä kohdekieltä puhuvan kanssa niin sujuvasti ja vaivattomasti ettei kumpikaan osapuoli koe vuorovaikutusta hankalaksi. (CoE 2018) Ruotsinkursseilla keskitytään opiskelijoiden alakohtaiseen kieleen, joskin yleiskielen ja rakenteiden opetusta on täytynyt lisätä heikentyneiden lähtötasotaitojen takia.

Toisen vuoden humanisteille ja yhteiskuntatieteiden opiskelijoille suunnatut ruotsin kielen kurssit ovat Turun yliopistossa viiden opintopisteen laajuisia; kontaktiopetusta on annettu 48 tuntia. Kielitaitonsa heikoksi kokeville opiskelijoille suositellaan kertauskursseja. Ruotsin kielen jäätyä pois pakollisina ylioppilaskirjoituksissa kirjoitettavien aineiden joukosta vuonna 2005 ruotsinkielentaito on laskenut tasaisesti, ja monen opiskelijan ruotsinkielentaito on toisen asteen jälkeen tasolla A2 (Juurakko-Paavola & Åberg 2018). Ruotsin kieltä on myös alettu kirjoittaa vähemmän ylioppilaskirjoituksissa. Taitotasosoa B1-B2 onkin ollut entistä vaikeampaa saavuttaa yliopiston ruotsinkursseilla aivan kuten Takala (2005, s. 293) on tutkimuksessaan ennustanut.

Ruotsin kielen kurssien kontaktiopetuksen jäädessä hyvin vähäiseksi on tullut tarpeen etsiä luokkahuoneen ulkopuolisia mahdollisuuksia harjoitella erityisesti puheen tuottamista. Turussa toimiva informaatiokeskus Luckan (http://abo.luckan.fi/) on tarjonnut siihen erinomaisen mahdollisuuden. Luckan jakaa informaatiota ja aineistoa Turun ruotsinkielisestä toiminnasta ja palveluista sekä antaa tilaisuuden puhua ruotsia muiden kielestä kiinnostuneiden kanssa. ”Pakkopulla” språkbadskaffe -keskusteluryhmä kokoontuu kerran viikossa kahtena eri ajankohtana ja osallistujille tarjotaan kahvia. Ryhmä mainostaa itseään vapaamuotoisella yhdessäololla ilman läksyjä, luentoja, kurssimaksua tai läsnäolopakkoa. Paitsi ruotsin kielen harjoittelua keskus tarjoaa myös paljon muuta toimintaa ruotsiksi, esim. elektroniikka- ja liikuntaneuvontaa, teatteria, lauluiltoja, kirjakerhoja ja alustettuja keskusteluiltoja.

Turun yliopiston Kielten ja viestinnän laitoksella kieliopintoja jatkaville humanistisen ja yhteiskuntatieteellisen tiedekunnan opiskelijoille Luckanin keskusteluryhmä on tarjonnut oivan tilaisuuden harjoitella kieltä luokkahuoneen ulkopuolella ilman opettajaa ja arviointia. Luckanissa opiskelijat ovat saaneet keskustella satunnaisten vierailijoiden kanssa ja heillä on ollut mahdollisuus löytää jatkumo kielenharjoittelulleen. Opettajalta opiskelijoiden Luckan-harjoittelu ei ole vaatinut erityisjärjestelyjä tai osallistumista. Ruotsinkurssilla olleiden opiskelijoiden tehtäväksi olen antanut tätä tutkimusta varten kirjoittaa raportin kokemuksistaan kielikylpykahvilassa.

4. Tutkimuksen tulokset

4.1 Aineisto ja sen analyysi

Formaali ja informaali oppiminen lähtökohtana halusin selvittää, miten opiskelijat kokevat kieliharjoittelun luokkahuoneen ulkopuolella – tässä tapauksessa kielikylpykahvilassa. Tutkimuskysymykseksi muotoutui: *Miten opiskelija kokee ruotsin kielen käytön Luckanin kielikylpykahvilassa.* Annoin opiskelijoille tehtäväksi käydä ainakin kerran Luckanin keskustelupiirissä kurssin aikana. Kokemuksistaan opiskelijoiden tuli kirjoittaa lyhyt raportti ruotsiksi, n 150–200 sanaa. Raportti sai olla vapaasti kirjoitettu, mutta siinä tuli keskittyä kertomaan ruotsin kielen käytöstä oppimiskokemuksena. Kerroin opiskelijoille, että käyttäisin raportteja tutkimustarkoitukseen, ja niiden käytön siihen tarkoitukseen voisi halutessaan kieltää. Kerroin myös, että raportteja ei palautettaisi korjattuina opiskelijoille eikä arvosteltaisi kielellisesti. Yksikään opiskelijoista ei

halunnut kieltää raporttinsa antamista tutkimustarkoituksiin ja tutkimusaspekti herätti opiskelijoissa kiinnostusta. Ensimmäiset tähän tutkimukseen keräämäni raportit ovat kevät- ja syyslukukaudelta 2016. Nämä ensimmäiset raportit opiskelijat saivat myös kirjoittaa suomeksi, mikäli ruotsi tuotti suuria vaikeuksia. Raportit ovat ainoastaan humanistiopiskelijoilta, jotka tuolloin olivat opetuksessani. Raportteja on yhteensä 68 kappaletta. Jatkoin raporttien keräämistä keväällä 2017, jolloin sain opetuksessani olevilta humanisteilta 16 raporttia ja yhteiskuntatieteilijöiltä 29 raporttia. Syksyltä 2017 humanistiraportteja kertyi 70 kappaletta. Keväällä 2018 suostuin opiskelijoiden pyyntöön vaihtoehtoisista mahdollisuuksista oppia luokkahuoneen ulkopuolella ja näin Luckan-raportteja kertyi humanisteilta vain 3 ja yhteiskuntatieteilijöiltä 4 kappaletta. Yhteensä Luckan-raportteja kertyi 190 kappaletta. Vaihtoehtoisina oppimistilaisuuksina opiskelijat käyttivät mm. Solsidan-elokuvaa, Turun tuomiokirkon ruotsinkielistä jumalanpalvelusta ja Turun linnan ruotsinkielistä turistikierrosta. Näitä raportteja kertyi humanisteilta ja yhteiskuntatieteilijöiltä yhteensä 34, ja niitä käytän tässä artikkelissa vain vähäisesti vertailumateriaalina luvussa 5.

Analyysissä raporttien ilmaisut pelkistettiin (ks. Tuomi & Sarajärvi 2018, s. 101), minkä jälkeen ne ryhmiteltiin yhtäläisten ilmaisujen joukoiksi. Samantyyliset ilmaisut alkoivat toistua, kun materiaalin lukeminen oli puolessavälissä. Jatkoin materiaalin lukemista ja analyysiä kuitenkin loppuun asti. (Ks. Eskola & Suoranta 2003, s. 63.) Analyysin jälkeen oli erotettavissa 4 erilaisten ilmaisujen luokkaa. Havainnollistan saamiani luokkia lukuisin raporteista poimituin esimerkein. Lainaukset ovat kaikki eri raporteista poimittuja ja siinä muodossa kuin opiskelijat ovat ne kirjoittaneet, jotta lukijalla olisi myös mahdollisuus tutustua opiskelijoiden ruotsinkielenkäyttöön laajasti.

4.2 Tunteiden kirjoa: hermostuttavaa, mukavaa, hauskaa, kiusallista, ihanaa…

Ylivoimaisesti eniten aineistosta löytyi ilmaisuja, jotka kuvasivat opiskelijoiden tunteita heidän osallistuessaan Luckanin keskusteluryhmään. Tunteiden on aivan viime aikoina myönnetty olevan erittäin tärkeä tekijä kielenoppimisessa, ja niiden merkittävyyden unohtaminen vieraan kielen opetuksessa on koettu yhdeksi kielenopetuksen pääongelmista (Dewaele 2015, s. 14; Swain 2013, s. 195–207; ks. myös Kaikkonen 2000, s. 55). On myös todettu, ettei oppimistilanteeseen liittyvien tunteiden tarvitse olla positiivisia, jotta oppimista tapahtuisi (Swain 2013, s. 195–207). Monet ruotsinkurssin opiskelijat keskustelivat vieraan ihmisen kanssa ruotsiksi ensimmäistä kertaa ja jännittivät tilannetta etukäteen tai kokivat sen ahdistavana. Opiskelijoiden jännittämisellä oli kuitenkin pikemminkin positiivinen vaikutus (ks. Waninge 2015, s. 198; van Lier 2000, s. 254): kaiken kaikkiaan keskustelutilanne koettiin miellyttävänä.

> En resa till Luckan kände spännande. Men javisst ska det känna spännande. Det är inte naturligt för en finländare att gå ha kafe och träffa med personer som han inte vet. Det ska känna specieltt pinsam om han mäste prata svenska.

> En diskussion var trevlig och ganska rolig i Luckan. Först var det ganska spännande att börjar prata ett okänt språk med okända människor.
>
> I Luckan var den svårtest saken gå in och sitta. I grupp var det trygg och trevligt. Jag tycker att erfarenheten var behaglig och hjälpt mig möta min räddes.
>
> Allt som allt var Luckan en skön erfarenhet. Jag fick värdefull träning med min artikulation.
>
> Luckan oli kokemuksena ihana. Mielestäni oli ihanaa päästä käyttämään sitä kielitaitoa, mitä olen opiskellut nyt 11 vuotta.

Monet kokivat keskusteluryhmän tuoneen uutta motivaatiota kielenopiskeluun, antaneen uutta intoa, rohkeutta, uskallusta ja energiaa jatkaa kielenopiskelua. Uusi innostus juonsi Luckanin hyvästä ilmapiiristä ja kokemuksen erilaisuudesta:

> Jag var inte överspänd och trivdes i Luckan. Jag är mer motiverad nu med svenska än tidigare. Jag ska lyssna svenska radioprogrammet och läsa tidningar. Jag är inte så osäker att prata svenska efter Luckan-erfarenhet. Jag ska förbättra min svenska med ny energi. Jag ska gå på Luckan med min kompis igen!
>
> Erfarenheten i Luckan var bra. Det var roligt att tala svenska, och ju mer jag talade, desto bättre kändes det. Personalen var vänlig och välkomnade och också gav ut gratis biljetter till Svenska dagen -festen på söndag. Luckan är också väl beläget i Åbo centrum och den var lätt att finna.
>
> Luckan oli myös sopivalla tavalla erilainen oppimismuoto ja se sopi itselleni hyvin. Luckanissa käyminen oli siis mielestäni antoisa kokemus ja se antoi erilaista motivaatiota ja intoa opiskeluun.

Seuraavissa lainauksissa tulee esiin se, miten tietoisen oppimisen mallin omaksunut opiskelija osaa muuntaa haluamansa tilanteen oppimistilanteeksi – tässä tapauksessa mahdollisesti pidempiaikaisemmaksi oppimistilanteeksi (ks. Takala 1992, s. 14; Williams & Burden 2007, s. 188). Opiskelijan on ollut mahdollista oppia oppimaan; taustalla on myös elinikäisen oppimisen malli:

> Vad lärde jag mig, frågar jag mig själv när jag går över torget hem till Mariegatan. Åtminstone prepositioner! Och att våga tala. Jag hade mycket trevlig kväll trots olyckbådande början. I trappan bestämmer jag: Jag har en ny hobby, varje onsdag kväll!
>
> Luckan kände lätt att träna på svenska. Alla människor var där på samma sak, att lära sig svenska. Där var också en bra val av svenskspråkig tidningar. Jag tittade på nätet att det finns många Luckan-kontorer utanför Åbo. Nu Jag vet åtminstone en ny, fri och utvecklande aktivitet att göra i större städer i Finland.

Täysin negatiivisia kokemuksia koko aineistosta löytyi vain kourallinen. Osa negatiivisesti suhtautuvista opiskelijoista piti parempana luokkahuoneopetusta, jonka yhteydessä annettiin selkeät ohjeet keskusteluaiheista. He olivat myös riippuvaisia tutusta

kurssitoverista. Virheiden pelko oli vahvasti kieliharjoittelupelon taustalla samoin kuin pelko käyttää vähäistä kielitaitoa ylipäänsä. Pelon vieraan kielen käyttämisessä on todettu vähenevän sitä mukaa kuin oppija edistyy; kielenkäyttöön tulee oppijan edistyessä myös lisää iloa (Dewaele 2015, s. 14). Viestinnällinen sujuvuus edellyttää kuitenkin kielen rakenteellisen hallinnan automatisoitumista (Kaikkonen & Kohonen 2000, s. 8; Järvinen 2014, s. 104). Juuri aroille oppijoille kannustus käyttää kieltä luokkahuoneen ulkopuolella on erityisen tärkeää, jotta he ymmärtäisivät myös kielen eri käyttömahdollisuudet (ks. Hildén 2000, s. 176–178).

> Hela tiden kände jag mig obekväm och jag skulle ha inte handlat med den här situationen om jag hade varit ensam…Generellt i livet känner jag inte bekväm med sådana situationer, och den här Luckan-besök var inte en avvikelse.
>
> Ymmärrän kyllä, että joillekin Luckanin tapainen ”vapaampi” kielenharjoittelu sopii hyvin, mutta itselleni ei. Minusta on paljon helpompaa harjoitella kieltä tilanteessa, jossa on selkeät ohjeet siitä, mitä tehdä ja mistä puhua.
>
> Puhumisen oppiminen motivoi lukemaan ja harjoittelemaan ruotsia, sillä huomasin sen ”tökkivän” aika paljon. Tunnilla tutun kaverin kanssa puhuminen on minusta mukavampaa, sillä vaikka aihe ei aivan täysin vapaa olekaan, minulla on rennompi olo ja pystyn kokeilemaan jonkin asian ilmaisua ja virheiden tekeminen ei pelota niin paljon, kuin vieraan kurssikaverin tai täysin vieraan ihmisen kanssa.

Seuraava kieliä opiskelevan opiskelijan toteamus on ainoa suomenruotsin ja ruotsinruotsin eroa käsittelevä ilmaisu:

> Verkligen trivdes jag mig inte eftersom jag inte vill lära mig finlandssvensk och jag hatar det. Jag tror att jag kunde dra nytta av Luckan i framtiden men jag är rädd för börja tala finlandssvensk. Det vill jag inte göra. Jag skulle vilja lära mig rikssvenskt uttal och vokabulär.

Lainauksessa esiin tuleva vastenmielisyys suomenruotsalaisuutta kohtaan on silmiinpistävä. Huolimatta siitä, että kielenopetuksessa huomioidaan nykyään aikaisempaa selvemmin kulttuurienvälinen oppiminen, on opiskelijan ymmärrys ja suvaitsevaisuus vähemmistöämme kohtaan jäänyt vähäiseksi (ks. Kaikkonen & Kohonen 2000, s. 9; ks. myös Karlsson-Fält 2010, s. 46). Vieraan kielen oppiminen on olennaisesti sidoksissa oppijan käsityksiin kielen kulttuurista samoin kuin hänen asenteisiinsa kohdekieltä puhuvia ihmisiä kohtaan (Williams & Burden 2007, s. 115–116). Tunteet voivat toimia välineenä kielenoppimistaitoja kehitettäessä (Imai 2010).

4.3 Kielenkäytön ja -oppimisen reflektointia

Kielenkäyttötilanteen autenttisuus antoi selvästikin opiskelijoille mahdollisuuden tiedostaa ja reflektoida sekä omaa että vierasta kielellistä käyttäytymistä ja kielenulkoista käyttäytymistä. (Ks. Kaikkonen 2000, s. 59.) Seuraavissa lainauksissa tulee esiin sekä kielenkäytön ja -oppimisen reflektointi että jokaisen oppijan yksilöllinen tapa oppia:

> Innan jag kom in, hade jag försökt att tänka på svenska hela vägen från hem till Luckan, så jag tror att det hjälpte mig också att lite förbereda mig, särskilt när jag först började att tala med de andra.
>
> Jag tycker att det är lättare att lära sig ord från en konversation än en lektion, därför att man kan komma ihåg konversationer bättre.
>
> Jag har aldrig tyckt om att tala mig själv och jag lär mig nya saker bäst när jag lyssnar och skriver, inte talar.
>
> Jag tycker att situation liknande den är en bra support för språkstudier, eftersom man ger erfarenhet av talande ett språk på normalt livssituationer istället att man måste stressade i klassrummet att få alla rätt.

Seuraava toteamus on mielenkiintoinen kertoessaan miten opiskelijatoverin hyvä kielenkäyttö vaikutti omaan motivaatioon positiivisesti. Luokkahuoneessa opiskelijatoverin osaamiseen ei tule kiinnitettyä huomiota:

> När jag hört annan student prata svenska väl på Luckan, fått jag motivation för att bli bättre så att jag kan använda svenska på ”äkta livet” också. När den samma situationen händer på klassrummet, påverkar det inte så mycket.

Kielenulkoisen käyttäytymisen pohdintaa tulee esiin seuraavassa ilmaisussa, jossa opiskelijan heikko kielitaito lienee estänyt hänen normaalin ystävällisen ja huomioivan käyttäytymisensä:

> Besöket på Luckan är de riktigt pinsama tre kvart till mig. Jag irriterar att jag verkades så ovänlig och glömmade mig bakom tekoppen även om de var so hänsynsfull.

4.4 Luokkahuoneen rajallisuus / Luckanin rajattomuus

Monet opiskelijat vertasivat raportissaan oppimista koulussa luokkahuoneen ulkopuolella oppimiseen. Erilaisuus rakentui monista seikoista: joillekin tilaisuus oli ensimmäinen puhetilanne luokkahuoneen ulkopuolella, toisille stressitön ilmapiiri ilman opettajaa ja valmista keskustelupohjaa oli ilonaihe. Kielikylpykahvila oli useimmille selvästikin sellainen oppimisympäristö, joka rohkaisi kielenkäyttöön. Se paransi itseluottamusta ja antoi rohkeutta kommunikoida vieraalla kielellä (ks. Williams & Burden 2007, s. 202).

> Jag anser att Luckan var en bra erfarenhet eftersom jag har talat svenska bara i skolan. Jag kunde öckså höra mycket svenska där så det är bra plats att lära sig mer svenska.
>
> Det var lätt att koppla av utanför klassrummet eftersom vi hade frihet att prata om vad som helst. En vanlig diskussion bredvid kaffebordet kan vara mer flexibel än en diskussion på en lektion. Man måste inte prata hela tiden: man får bara lyssna på andra om man vill.

> Tilaisuus avasi silmiä myös siinä mielessä, että oivalsin ruotsin puhumisen olevan helppoa paineettomassa ympäristössä. Otankin tästä lähin tavoitteekseni ”rattilukon” poistamisen arkitilanteiden kielenkäyttötilanteissa. Vanhat herratkaan eivät olleet millään tavalla pahoillaan pienistä virheistäni mietintätauoistani. Jos viesti menee perille, on tärkein jo kunnossa. Suomalaiseen opetussuunnitelmaan pitäisi ehdottomasti sisältyä jonkinlainen rohkaisuun ja terveeseen itseluottamukseen tähtäävä kokonaisuus.

> Vi talade med varandra in en liten grupp om allt som vi skulle tänka om, vilket var olikt från ett klassrum: vi hade inte på förhand bestämda saker som vi måste tala om men vi fått själv välja vad vi ville tala om. Det var kul. Vi var också erbjudits gratis biljetter till Åbo Svenska Teaterns Peter Pan – skådespel, vilket var underbart.

Jotkut kuvailivat luokkahuoneopetusta passivoivaksi, kun taas sen ulkopuoleinen oppiminen vaatii todellista keskittymistä:

> Lektionen på klassrummet består mera av grammatik och där man kan vara ganska passiv men stunden i Luckan kräver att du måst våga tala med okända människor och improvisera om du inte vet något ord.

> Tilanne oli myös opetukseen verrattuna erilainen, sillä läsnäolo keskustelutilanteessa vaatii mielestäni jossain määrin enemmän keskittymistä, kuin opetus luokkahuoneessa… Puhuminen on mielestäni erittäin tehokas tapa vahvistaa kielitaitoa, joten tämän kaltaiset vierailut ovat tehokkaita opetuksen muotoja.

Osalle opiskelijoita autenttinen tilanne antoi aihetta pohtia oman kielenkäytön rajoitteita. Heikoiksi jääneiden kielioppitaitojen sekä suppeaksi jääneen sanaston koetaan aiheuttavan vaikeuksia. Kieliopin opetuksen on todettu nopeuttavan nuorten ja aikuisten kielen oppimista (ks. Jaakkola 2000, s. 150). Sanaston oppiminen yliopistokurssilla on alkanut vaikeutua kunkin opiskelijan käyttäessä henkilökohtaista tietokonettaan: keskustelutilanteessa näytöllä on yleensä helppokäyttöinen sanakirja, joka kyllä antaa kaivatun sanan senhetkiseen käyttöön, mutta mieleen painaminen vaatisi sitoutumista harjoitteluun (vrt. Jaakkola 2000, s. 150).

> I Luckan jag lärde att jag måste läsa svenska mer och lära bättre grammatik. Jag också märkte att jag måste praktikera mera vokabulär.

Monet opiskelijoista löysivät autenttisessa tilanteessa mahdollisuuden käyttää kieltä kokonaisvaltaisesti: ajatukset eivät enää olleet kielen yksittäisissä osa-alueissa, vaan ymmärtämiseksi tulemisessa (ks. van Lier 2000, s. 255; Kramsch 1993, s. 4–5).

> Man kanske inte alltid tänker på det men i vardagliga situationer behöver man inte stressa över grammatik och vokabulär. Alla ju blir alltid förstådda till slut. I skola, där man vill lära sig svenska är det naturligtvis viktigt för läraren att korrigera eventuella grammatiska fel. Men i situationen som händer utanför klassrummet finns det kanske inte samma roller.

Rajoittavana tekijänä koetaan luokkahuoneessa myös toisen opiskelijan huono kielitaito tai huono motivaatio:

> På lektionerna vid universitetet är det ofta tråkigt eftersom de flesta studeranden kan inte prata svenska och de tycker inte att använda den.

> Det var import att pensionerna hade motivet att lyssna på mig. Studenterna har inte motivet för svenska ibland på universitetet.

Useille opiskelijoille kahvilan erilaiset ihmiset olivat miellyttävä lisä opiskeluympäristössä; myös keskustelunaiheet olivat erilaisia kuin luokkahuoneessa. Autenttisessa tilanteessa opiskelija joutuu huomioimaan kokonaisvaltaisesti myös keskustelukumppaninsa, mikä tuntuu tehostavan oppimista (ks. van Lier 2000, s. 254; Kramsch 1993, s. 9–11).

> Jag märkte att en annan grupp påverkade min motivation och mitt lärande. Att prata med en okänd främling gjorde att jag verkligen vill försöka mitt bästa, och jag förstod också hur viktigt det är att lyssna noga på varandra för att bättre förstå det.

> Det var också kul att få diskutera med olika människor, därför att vi pratade om saker man kanske inte diskuterar om med andra universitetsstudenterna och i klassrummet.

> Luckan gruppen hade olika språktalare lika som i universitets gruppen. Där var blivande svenska lärare, hon kände svenska så bra att det underlättade en gemensam diskussion.

4.5 Pakkoruotsia vai riemuruotsia?

Kurssilla olleille opiskelijoille käynti kielikylpykahvilassa oli pakollista. Pakollisuus ei kuitenkaan tullut esiin raporteissa negatiivisena asiana vaan positiivisena: pakko oli monen mielestä hyvä asia. Näyttäisikin siltä, että opiskelijat saivat luokkahuoneen ulkopuolella aidon syyn kommunikoida, koska vastapuolena oli joku muu kuin opiskelijatoveri. Kurssilaisten kanssa keskustelu luokkaympäristössä saattaa jäädä pelkästään rutiininomaiseksi tehtävän suorittamiseksi. Pakkoruotsista tulikin Luckanissa monelle riemuruotsia, joksi suomenruotsalainen vähemmistö on ns. pakkoruotsin nimennyt.

> Det var jätte bra att vi måste gick i Luckan, därför att jag förstå att talar svenska är bäst övningen.

> Jag vet att upprätthållning av språkkunskap kräver sådana här aktiviteter, och det skulle vara väldigt nyttigt att prata mer. Tyvärr är det bara för lätt att glömma när det är inte längre obligatoriskt.

> Jag var också mera motiverade att försök mitt bästa, eftersom jag inte ville skända mig. Jag också lärde mig många nya ord från svenskspråkiga människor. Allt som allt det var jätte bra att vi måste gå till Lucan.

> Det var ett bra idé att besöka Luckan och använda svenska i en realistisk och olik miljö.
>
> Kaiken kaikkiaan Luckan oli minusta yllättävän miellyttävä kokemus. Voisin jopa harkita käyväni siellä joskus uudestaan. Kun oli niin sanotusti pakko käyttää kieltä, se oli jotenkin paljon luontevampaa kuin teennäisissä luokkakeskusteluissa.

5. Solsidan, Turun tuomiokirkko, Turun linna etc.

Keväällä 2018 opiskelijat saivat mahdollisuuden käydä myös jossakin vaihtoehtoisessa ruotsinkielisessä oppimisympäristössä ja Luckan-raportteja kertyi vain muutama. Sen sijaan raportteja tuli paljon Solsidan-elokuvasta, Turun tuomiokirkon ruotsinkielisestä jumalanpalveluksesta ja Turun linnan ruotsinkielisestä ritarikierroksesta. Yksi opiskelija oli omaa alaansa käsittelevällä Studia Generalia -luennolla Åbo Akademissa, pari tutustui ruotsinkieliseen akateemiseen ”spexiin” (”Akademiska Spexet vid Åbo Akademi” on humoristista ylioppilasteatteria, jonka näyttämökohtauksiin yleisö voi vaikuttaa katsomosta huutelemalla. https://spex.abo.fi/spex). Näistä raporteista voi erottaa paljon samoja teemoja kuin Luckan-raporteista. Yksi selkeä ero Luckan-raportteihin nähden oli kuitenkin tunteiden kirjon niukkuus, mikä saattoi johtua siitä, ettei opiskelijan tarvinnut jännittää omaa puhumistaan. Myös ajatukset omasta suullisesta kielenkäytöstä puuttuivat. Suurin osa näistä kielenkäyttötilanteista osoittautui passiivista kielitaitoa tukevaksi; opiskelijat saivat testata puheenymmärtämistaitojaan. Jotakuinkin kaikki opiskelijat kokivat informaalin oppimistilanteen hyvin positiivisena. Opiskelijoiden oma kielenkäyttö rajoittui pääasiallisesti puheen ymmärtämiseen. Erilaisen ympäristön kokeminen tuli esiin monissa toteamuksissa:

> Stor och vid domkyrka fick röster att eka runt salen så det var väldig svår att förstå vad prästen sade. I alla fall fattade jag en lejonpart av hans liturgi. (Turun tuomiokirkko)
>
> Några gånger var det också lite förvirrande när publiken började att skratte och jag hade ingen anign varför. (Akademiska spexet vid Åbo Akademi)
>
> Ytterligare det var ganska svårt att förstå barn, eftersom barn pratade oklar. (Turun linna)

Näissäkin raporteissa tuli esiin paljon vertailuja luokkahuoneoppimisen ja luokan ulkopuoleisen oppimisen välillä. Luokkahuoneen ulkopuoleinen ympäristö koettiin eri tavoin vapaammaksi, koska monet opiskelijatoverit ja kielitaidon kontrollointi puuttuivat. Toisaalta kokemuksissa nousee esiin arvostus luokkahuoneopetuksen aktivoivaa toimintaa ja oppimisympäristön tarkoituksenmukaisuutta kohtaan.

> Inlärning utan lärare var i något sätt mer fri (friare?) när man inte blir medveten om sina misstag. Även om jag förstår att det är viktigt att få feedback om misstagen känns det olika när man inte kan jämföra sig själv till andra studerande. I alla fall var filmet

> en roligt erfarenhet och högklassig kultur upplevelse, även om alla karaktärer pratade svenska. (Solsidan)

> Det var ganska svårt att förstå prästens språk i ekonande kyrkan. Det fanns många obekanta ord i mässan så att min förståelse led. När jag jämförar inlärning i klassrummet med inlärning i kyrkan är det mycket olik. I klassrummet kan man vara mer interaktiv och fysiska omständigheter är bättre för inlärning. Jag anser att de var endå roligt att lära sig nytt hus fri miljö med goda vänner. (Turun tuomiokirkko)

> Sådant studiet av svenska språken är trevlig och lite mera konkretisk än studiet i klassen. Men i sådant situationer mans egna deltagande är lite mindre eftersom man behöver inte att talas så mycket. Jag tycker att det är nyttigare och effektivare att studera språken i klassen eftersom där måste vara mera aktiv och ordet vad man använder är bekantare. (Turun linna)

Kaikista raporteista vain yhdessä opiskelija kuvailee hakeutumistaan oppimisympäristöön, jossa pääsi kuulemaan omaa alaansa käsittelevää kieltä:

> Jag tycker att lära mig terminologin på det här sättet var ganska effektiv: jag kunde inte koncentrera mig på grammatik (jag hade ingen tid!) utan jag bara plockade de logopediska begreppen som jag ville veta på svenska. (Studia generalia vid Åbo Akademi)

Vaikka puhuminen saattoi tuottaa opiskelijoille suuria vaikeuksia, se oli ilmeisesti monille kuitenkin helpompaa kuin kirjoittaminen. Seuraava lainaus kuvaa erään elokuvissa käyneen, myönteisesti informaaliin oppimisympäristöön suhtautuneen opiskelijan kirjoittamistuskaa:

> Det allra tråkigaste saken i den här uppgiften var att skriva den här uppsatsen. Det var svårt att hitta på vad jag kunde skriva. Den här var mycket svårtare skriva än till exempel ett referat. Lyckligtvis är 150 ord inte så mycket. (Solsidan)

Vaikka opettajana halusinkin antaa opiskelijoille ensisijaisesti tilaisuuden käyttää aktiivisesti kielivarantoaan informaalissa oppimisympäristössä, totesin kaikki raportit luettuani myös kuuntelua kehittävät oppimisympäristöt todella arvokkaiksi paikoiksi kielen kehittämisen kannalta. Ruotsin kielen nykyisen osaamistason ja opiskelijoiden kielenkäyttöarkuuden huomioiden on tärkeää löytää eritasoisille ja oppimistavoiltaan erilaisille opiskelijoille monenlaisia oppimisympäristöjä.

6. Tulosten pohdintaa

Luckan-raporttien antia pohtiessa huomio kiinnittyy ennen kaikkea tunteiden vahvaan läsnäoloon kielenoppimisessa. Korkeakoulussa, oppijoiden ollessa aikuisopiskelijoita, tunneaspekti unohtuu helposti opettajan pitäessä itsestään selvänä opiskelijan kykyä toimia itseohjautuvasti ja käyttää kieltä rohkeasti. Olisi kuitenkin tiedostettava, että

suurimmalla osalla tämän päivän ruotsinopiskelijoita on jo lähtökohtaisesti aikaisempia opiskelijapolvia huomattavasti heikompi ruotsinkielentaito, ja näin ollen kynnys puhua kieltä on korkealla. Monille pakollinen ruotsinopiskelu korkeakoulussa on ahdistavaa ja jää usein tutkinnon viimeiseksi suoritusosioksi. Kynnystä puhua ruotsia voidaan varmasti madaltaa ohjaamalla opiskelijoita myös korkeakoulutasolla entistä aktiivisemmin paitsi sähköisten oppimateriaalien pariin, myös autenttisiin informaaleihin oppimisympäristöihin. Parasta kuitenkin olisi aikainen aktiivinen puheentuottaminen jo koulussa, sillä korkeakoulussa varsinkin pakollinen kieli joutuu kilpailemaan monen muun tärkeän asian kanssa eikä kieliharjoittelulle enää jää aikaa.

Toisena varteenotettavana seikkana, joka raporteissa tulee esiin tunteiden ohella, on opiskelijoiden positiivisuus erilaisia oppimisympäristöjä kohtaan. Vapaus – tai pakko – uskaltaa käyttää kieltä tuottaa useimmissa opiskelijoissa iloa. Ruotsiakin voi puhua muiden kuin opettajan ja kurssitovereiden kanssa, ja samalla voi oppia omasta ja muiden kielenkäytöstä. Myös informaalin oppimistilanteen voi oppia hyödyntämään oppimistilanteena; voi oppia ohjaamaan sekä omaa opiskeluaan että ymmärtämään elinikäisen oppimisen mahdollisuuden.

Ilahduttavana seikkana raporteissa tulee esiin suhtautuminen ruotsin kieleen kuin mihin tahansa muuhunkin opittavaan kieleen. On tietysti huomattava, että tämän tutkimuksen kohteina olivat humanistit ja yhteiskuntatieteilijät, jotka suhtautuvat ruotsin kielen oppimiseen yleensä positiivisesti. Suomenruotsalaiseen kulttuuriin ja ruotsinkielen eri variantteihin tutustuttaminen jo koulutasolla olisi tärkeää. Pohjanmaan murteet ja Ruotsin maahanmuuttajien ruotsinkieli voisivat olla mielenkiintoisia tutustumiskohteita.

Kirjallisuus

Akademiska Spexet vid Åbo Akademi. Saatavilla: https://spex.abo.fi/spex (12.11.2018)

Banks, J., Ball, P., Gordon, E., Gjutierrez, K., Heath, S., Lee, C., Lee, Y., Mahiri, J., Nasir, N., Valdes, G. & M. Zhou 2007. *Learning in and out of school in diverse environments. Life-long, life-wide, life-deep.* Saatavilla: http://depts.washington.edu/centerme/ home.htlm (9.4.2018)

Benson, P. (2013). Teaching and researching autonomy in language learning. Applied Linguistics in Action. Hoboken: Routledge.

Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Companion Volume with New Descriptors, 223.

Dewaele, J-M. (2015). The Language Teacher 39.3. JALT Conference Article. Saatavilla: http://jalt-publications.org/tlt (29.4.2018)

Dewey, J. (1948). Demokrati och uppfostran. Natur och kultur. Stockholm. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors. Saatavilla: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989 (12.11.2018)

Eskola, J. & J. Suoranta. (2003). Johdatus laadulliseen tutkimukseen. Jyväskylä: Gummerus.

Hager, P. & J. Halliday. (2009). Recovering Informal Learning. Wisdom, Judgement and Community. Springer Science + Business Media B.V.

Hildén, R. (2000). Vieraan kielen puhuminen ja sen harjoittelu. Teoksessa P. Kaikkonen & V. Kohonen (toim.) *Minne menet, kielikasvatus?* Jyväskylä, 169–180.

Huhta, A. (1993). Teorioita kielitaidosta - Onko niistä hyötyä testaukselle? Teoksessa S. Takala (toim.) *Suullinen kielitaito ja sen arviointi*. Julkaisusarja B: Teoriaa ja käytäntöä 77. Kasvatustieteiden tutkimuslaitos, Jyväskylän yliopisto, 77-142.

Jaakkola, H. (2000). Kielitiedosta kielitaitoon. Teoksessa P. Kaikkonen & V. Kohonen (toim.) *Minne menet, kielikasvatus?* Jyväskylä, 145–156.

Imai, Y. (2010). New Insights from Collaborative Learning for an EFL Classroom. *The Modern Language Journal*, 94(2), 278–292.

Juurakko-Paavola, T. & A-M. Åberg. (2018). Ruotsin kielen osaamisvaatimuksista vapauttaminen korkeakouluissa. Kieli, koulutus ja yhteiskunta, 9(1). Saatavilla: https://www.kieliverkosto.fi/fi/journals/kieli-koulutus-ja-yhteiskunta-maaliskuu-2018/ruotsin-kielen-osaamisvaatimuksista-vapauttaminen-korkeakouluissa (12.11.2018)

Järvinen, H-M. (2014). Kielen opettamisen menetelmiä. Teoksessa P. Pietilä & P. Lintula (toim.) *Kuinka kieltä opitaan. Opas vieraan kielen opettajalle ja opiskelijalle.* Gaudeamus Oy, 89–113.

Kaikkonen, P. (2000). Autenttisuus ja sen merkitys kulttuurienvälisessä vieraan kielen opetuksessa. Teoksessa P. Kaikkonen & V. Kohonen (toim.) *Minne menet, kielikasvatus?* Jyväskylä, 49–61.

Kaikkonen, P. & V. Kohonen. (2000). Minne menet kielikasvatus? Teoksessa P. Kaikkonen & V. Kohonen (toim.) *Minne menet, kielikasvatus?* Jyväskylä, 7–10.

Karlsson-Fält, C. (2010). Kielikeskuksissa toimivien kieltenopettajien käsityksiä ja kokemuksia massayliopiston luomista haasteista. Turun yliopisto. Turun yliopiston julkaisuja 303.

Karlsson-Fält, C. & M. Maijala. (2007). Kieltenopettaja ja kielen oppiminen – mitä opiskelijat arvostavat opettajassa? Teoksessa O-P. Salo, T. Nikula & P. Kalaja (toim.) *Kieli oppimisessa - Language in Learning*. AfinLA:n vuosikirja 2007. Suomen soveltavan kielitieteen yhdistyksen julkaisuja 65. Jyväskylä, 331–350.

Kramsch, C. (1993). *Context and Culture in Language Teaching*. Oxford University Press.

Krashen, S. D. (1981). *Second Language Acquisition and Second Language Learning*. University of Southern California.

Kuokkanen-Kekki, M. & C. Niedling. (2011). Learner autonomy and intercultural exchange in language learning: ALICE as institutional Tandem. Teoksessa K.K. Pitkänen, J. Jokinen, S. Karjalainen, L. Karlsson, T. Lehtonen, M. Matilainen, C. Niedling & R. Siddall (toim.) *Out-of-Classroom Language Learning*. Language Centre Publications 2. Helsinki University Printing House, 39–54.

Richards, J. C. (2015). The Changing Face of Language Learning: Learning Beyond the Classroom. *RELC Journal* 2015, Vol. 46(I) 5–22.

Siekkinen, H. (2017). Työelämä, täältä tullaan – kielitaitoisina ja itseohjautuvina. Tutkimus opiskelijoista monimuotoisesti toteutetulla englannin opintojaksolla Jyväskylän ammattikorkeakoulussa. Jyväskylän yliopisto.

Singh, M. (2015). *Global Perspectives on Recognising Non-formal and Informal Learning. Why Recognition Matters.* Springer.

Siurala, L. (2002). *Can youth make a difference? Youth policy facing diversity and change.* Council of Europe.

Swain, M. (2013). The inseparability of cognition and emotion in second language learning. Plenary Speeches. *Language Teaching*. (2013), 46.2, 195–207. Cambridge University Press 2011. First published online 28 November 2011.

Takala, S. (1992). *Virikkeitä uutta kokeilevaan koulutyöhön*. Kasvatustieteiden tutkimuslaitos. Jyväskylän yliopisto. Kirjapaino Oy Sisäsuomi.

Takala, S. (1997). Kielen oppiminen taitotasolta toiselle. Teoksessa M. Tarnanen, S. Takala, A. Määttä, A. Huhta, & B. von Bonsdorff (toim.) *Yleiset kielitutkinnot ja kieltenopetus.* Helsinki: Hakapaino Oy, 85–95.

Takala, S. (2005). Reflections on the development and impact of language education in Finland. Teoksessa J. Smeds, K. Sarmavuori, E. Laakkonen & R. de Cillia (toim.)

Multicultural Communities, Multilingual Practice. Monikulttuuriset yhteisöt, monikielinen käytäntö. Festschrift für Annikki Koskensalo zum 60. Geburtstag. Turun yliopisto, 285–296.

Tuomi, J. & A. Sarajärvi. (2018). *Laadullinen tutkimus ja sisällönanalyysi*. Kustannusosakeyhtiö Tammi.

van Lier, L. (2000). From input to affordance: Social-interactive learning from an ecological perspective. Teoksessa J. P. Lantolf (toim.) *Sociocultural Theory and Second Language Learning*. Oxford: University Press, 245–259.

Waninge, F. (2015). Motivation, Emotion and Cognition: Attractor States in the Classroom. Teoksessa Z. Dörnyei, P. D. MacIntyre and A. Henry (toim.) *Motivational Dynamics in Language Learning*. Short Run Press Ltd. Great Britain, 195–213.

Williams, M. & R. L. Burden. (2007). *Psychology for Language Teachers. A social constructivist approach.* Cambridge: Cambridge University Press.

Dr. Sauli Takala, a coach who made a difference

Olga Lankina

St Petersburg State University

Professor emeritus Sauli Takala, a distinguished scholar and a man of many virtues, has definitely left a significant mark on people who knew him professionally and personally. In this short entry I will write about Dr. Takala as a lecturer and a coach who always maintained high standards of work ethics.

I was happy to get to know Dr. Takala in 2013 when he and his colleagues *Dr*. Neus *Figueras Casanovas* and *Dr*. *Norman Verhelst* arrived in St Petersburg to coach a CEFR linking project. The test in question was the exit test in General Academic English for Bachelor level students of St Petersburg State University developed by the team of local raters led by Dr. Elena Prokhorova. Within this project the coaches, including Dr. Takala, provided invaluable support and guidance at all the stages of this complex endeavour.

In June 2014 Dr. Takala gave a presentation on the specifics of the Language in Use paper. It was highly informative and helpful and covered a vast array of topics related to grammar and vocabulary in relation to language assessment. Delivered in a professional and calm manner, so typical of Dr. Takala, this talk is remembered by my colleagues as an example of a true academic approach to the subject. The slides that he used for this presentation and then shared kindly with the St Petersburg team of raters revealed Dr. Takala's devotion to perfection: every aspect of the topic was covered scrupulously. This presentation as well as the presentations of the other coaches was a perfect introduction to the Standard Setting for the receptive skills.

The Standard Setting for the productive skills was marked by another memorable presentation made by Dr. Takala in November 2014. This time it was on Writing, the subject which was of particular interest to Dr. Takala. Dr. Takals's talk spanned all possible issues of Writing from its origin to Writing in the CEFR. The audience could not but feel Dr. Takala's passion for this subject which manifested itself in a comprehensive and complete exposure of the listeners to the topic.

As the project went on, the St Petersburg raters approached Dr. Takala on many occasions: those were the issues of obtaining some training materials or seeking Dr. Takala's advice. He was always generous in sharing and very quick to respond and offer practical and effective solutions. While providing help or giving feedback,

Dr. Takala was professional but tactful and considerate which inspired people who corresponded with him.

On top of that Dr. Takala was a man of many interests. I remember Dr. Takala making witty remarks and giving detailed comments on pieces of art during the tour of the Hermitage. He was also a fun-loving man and had a good sense of humour and his eyes would sparkle when people shared a good joke.

I am sure that we all have learnt good lessons from Dr. Takala and we owe many of our successes to the knowledge and expertise that Dr. Takala shared with us.

Land skall med lag byggas

Christer Laurén
Vasa Universitet
(University of Vaasa)

Abstract

'Land shall be built on law'

The aim of this article is to give an idea of the earliest justice in Finland-Sweden with an emphasis on the practical application of law and tradition. Witnessing at court required men regarded as trustworthy by the court. It was, for example, enough to certify that they did not think that the defendent could have offended against the law. Judicial penalty for crime or violation of the law was in many cases a physical punishment, for example, cutting off the arm of the one who was guilty of theft. One thing we have to remember, though, is that the population was very small and the few inhabitants of the small villages and towns knew each other fairly well during the medieval period and the first centuries of the modern age.

”I have a dream… ” som Martin Luther King började det tal som ryckte med sig alla som hörde honom; till och med texten efteråt får oss att drömma den goda drömmen. Men det är också möjligt att använda retoriken för negativa, onda reaktioner. Vältalighet, konsten att formulera sig väl, är ingenting värd om man inte har ett budskap som är angeläget för en själv och för ens åhörare och verkligen vet att analysera det på ett insiktsfullt sätt.

Vi har väl alla upplevt att vi inte har fått rätt även om vi varit övertygade om att vi hade rätt. På kvällen går vi igenom det som hänt och vad vi kunde ha sagt bättre och mer övertygande. Vältalighet, talekonst, retorik, är ett sätt att övertyga och man kan vara mer eller mindre skicklig vältalare eller skribent. Det finns ingenting som totalt saknar retoriska kvaliteter. Det pågår alltid samtal som vi deltar i och vi påverkar alltid den/dem vi kommunicerar med. I en demokrati har vi alla en röst och vi har alla skyldighet att använda den.

Min favorit bland vältalarna i Rom för tvåtusen år sedan är Cicero, advokaten, som lämnat efter sig 900 brev och 50 tal. Ett av talen är försvarstalet för Milo. Cicero hade nämligen misslyckats att försvara sin klient och han var en av oss som låg och vände sig i sängen på natten, steg upp och började den nya dagen med ett förbättrat tal, ett av de bästa han någonsin hållit. Men Milo var redan dömd. Cicero kunde bara låta

talet ingå i sina bokrullar om vältalighet. Ciceros böcker är också i vår tid värda att läsas som läroböcker.

* * *

Mitt syfte med denna text är att ge en historisk tillbakablick främst ur vårt samhälles synvinkel. Det blir ett antal nedslag i händelser och seder i en tusenårig historia om rättsväsendet och dess funktion. När människor flyttar samman, bildar samhällen, blir det också behov av att reglera umgänget mellan dem på ett vettigt sätt. Det måste finnas ett sätt som är konsekvent och som upplevs som rättvist. Man brukar säga att lagen och dess tillämpning växer fram ur människors rättsmedvetande. Detta romantiserande sätt att se på lag justerades i mitten av 1970-talet, när det visades att det fanns inflytande från centraleuropeiskt tänkande och att de medeltida lagarna visade spår av olika maktkonstellationers intressen. På grund av områdets ålder är språkbruket präglat av en viss konservatism och därmed ofta stelt och tungt. Men det har också estetiska kvaliteter, det kan upplevas som vackert.

Vi skall tänka oss tillbaka i tiden till ett samhälle som saknade tryckkonst och till och med skrifttraditioner. Där måste man lita på sitt minne. Kulturen är muntlig och minnet behöver hjälpmedel som rytm och rim och konkreta berättelser. De första bevarade skrivna lagsamlingarna på svenska (landskapslagar på fornsvenska) har i sitt språk spår av en muntlig kultur. Där finns t.ex. berättelser som den om vilka skyldigheter man har när man är ute till havs och tar i land på en ö, och man hör knackningar i en stenhäll och rop på hjälp. Man är då skyldig att hjälpa den som är instängd, av sina fiender eller av en olycka, han kan ha varit skendöd. Från sådana konkreta berättelser skulle man avleda tolkningen av andra, liknande berättelser.

Man erinras också om att det inte alltid är människor som bär skuld. Det kan enligt en landskapslag också vara en tupp som skrämd och flaxande flyger upp när en gäst träder in. Tuppen satt på en yxa som var inkilad mellan stockar ovanför dörren och denna yxa föll ner och dödade gästen. Tuppen blev då den mannens bane.

Det fanns under landskapslagarnas tid en regel som sade att den som kom in under sotad ås hade rätt att bli behandlad som en gäst i huset. Den sotade åsen var taket i huset på den tiden när man hade eld i mitten i huset och röken sökte sig uppåt mot ett hål i taket. På den tiden fanns det inte andra möjligheter för inkvartering än hos vanliga familjer.

Sådana berättelser var dessutom formulerade med rytm och rim. ”Gånge hatt till, huva från” innehåller förutom allitterationen en motsats i form av till och från. Det korta uttrycket innebär att mannen har förtur till arv före kvinnan. Emellertid tilltar allitterationerna med tiden, inte som man romantiserande trott att de avtagit.

De äldsta bevarade svenska landskapslagarna Äldre Västgötalagen och Östgötalagen är från 1200-talet. Upplandslagen är stadfäst av konungen och har därför också det bästa anseendet. Man har kunnat visa att Upplandslagens författare (bättre: nerskrivare) var juridiskt bildade män som kände till den romerska rätten.

Helsingelagen gällde för Österlandet, vilket var ett område i det nuvarande sydvästra Finland. Helsinge berättar något om varifrån utflyttare till Finland kommit

under vissa tider. Andra bevarade landskapslagar är Smålandslagen, Västmannalagen, Dalalagen och Södermannalagen. Alla dessa lagar, bortsett från Upplandslagen, är privata uppteckningar. Dalalagen anses mest ursprunglig. Gutalagen som gäller för Gotland hör inte till den svenska utan till den danska traditionen. Skånelagen från 1200 är likaså en gammeldansk lag.

* * *

Därmed kan vi se på en estetiskt tilltalande text, nämligen Skånske Lov som har en praktutskrift från c. 1300 med runor på pergament, Codex runicus. Den avslutas med en text om gränsdragningen mellan Sverige och Danmark. Därpå följer en dikt med noter som har använts i Danmarks radio som paussignal. Det överraskar en modern läsare av juridiska texter. Dikten är vacker liksom försök som finns att återge melodin. Ofta har man sett dikten som en kärleksdikt nerskriven allra sist av en trött avskrivare efter en lagtext. Men handskriften är en lyxutgåva som man inte slarvat med. Dessutom är den hand som skrivit texten samma hand som har skrivit de föregående sidorna. Troligen är handskriften avsedd att vara en gåva till en högt uppsatt person. På gammeldansk lyder dikten på följande sätt:

Drømde mik en drøm i nat um

silke ok ærlik pæl

(Jag hade en dröm i natt

om silke och finaste päls; alternativt:

Jag drömde en dröm i natt

om rättvisa och ärlighet)

Den romantiserande tolkningen har varit att det är fråga om en kärleksdikt. Vad annat kunde passa bättre in på den skrivare som blivit färdig med en lagtext än att drömma om hans älskade. Visserligen kan man från metafor ta språnget till metarens metafor. Kanske freden bara är en dröm som man kan drömma om, som det heter på gammeldansk.

Men den seriösa tolkningen är att man äntligen efter krig kommit överens om gränsdragningen mellan Sverige och Danmark. Den som upplevt krig och längtar efter fred förstår att man behöver starka metaforer.

* * *

I äldre germansk rätt var bevisgången annorlunda än i vår tid i Finland. Det vanligaste sättet var svarandens värjemålsed förstärkt med ed (dulsed) av edshjälpare eller edgärdsmän.

Under medeltiden och i början av nya tiden undanträngdes detta sätt att bevisa. I Sverige bibehölls edgärdsmannatraditionen och fick en viktig roll i landskapslagarna (från 1500- och 1600-talen). Den upphävdes först genom en kunglig förordning 1695. Den innebar att en åtalad mot vilken indiciebevis fanns kunde hänvisa till släktingars, vänners och grannars vittnesmål om att de var övertygade om att den åtalade inte hade utfört brottet.

Det låter märkligt för oss att man kunde fälla en åtalad på så, som vi ser det, lösa grunder. Vi måste då komma ihåg att samhällena var små och man kände varandra rätt väl och man hade en vördnad för eden av religiös art. Den som begick mened hade att vänta både straff i evigheten och straff i denna världen.

* * *

Ett fall i Gamlakarleby gällde ett inbrott i en strandbod som i slutet av 1600-talet begåtts av Johan Larsson. Larsson uppgav att hans medbrottsling var hans svåger borgaren Thomas Jacobsson Tast - som nekade. Medan Johan under rannsakningstiden hölls fängslad fick Tast som var en förmögen borgare vara på fri fot. Eftersom Tast var förmögen antog man att han inte skulle rymma. Men det samlades indicier som pekade på att Tast var skyldig. Rätten beslöt att Tast inom tre veckor skulle svära sig fri själv tolfte. Inom den tiden infann han sig inför rätten med sina edgärdsmän, fem bönder från Lochteå, fyra bönder från Gamlakarleby socken och sina två svågrar från staden. Rätten betonade för dem att de skulle noga tänka sig före, att det kunde finnas risk för mened. Några av bönderna blev osäkra och vägrade gå ed.

Några dagar senare kom Tast med nya vittnen, svågrar, svägerskor och syskonbarn. Igen var det en av bönderna som blev osäker. Han hade av sina sockenbor blivit varnad och vägrade gå ed. Motparten förklarade då att Tasts släktingar var jäviga p g a släktkärlek och att bara stadsbor som kände den åtalade kunde begå eden. Rätten blev då osäker och vände sig till hovrätten med förfrågan om släktingar, man och hustru, släkt och svågrar kunde gå ed i ett sådant fall.

Ett helt år senare kom hovrättens beslut som betydde att fyra Lochteåbönder och Tasts dräng fick begå eden, Tast själv sjätte. Tast blev frikänd även om allt tydde på att han var skyldig. Detta hände i slutet av 1600-talet.

* * *

Det finns flera rättsfall som gällde slagsmål och skränande i staden. Det var dock en förmildrande omständighet om en av parterna hänvisade till att han var berusad och därför inte var vid sina sinnens fulla bruk. Den nyktre förutsattes vara försiktig och därför visa behärskning.

Tortyr hade på det europeiska fastlandet kommit i bruk genom katolska kyrkan. Tortyr för att framtvinga bekännelser var inte vanligt i Sverige förrän i slutet av 1600-talet – och grov tortyr främst vid politiska mål. I slutet av seklet förbjöds de av Karl XI, den konung som grundade Vasa stad.

Böter förekom mycket ofta. I slutet av 1600-talet förekom gatlopp. Nakna till midjan skulle de dömda springa mellan två led av 100 män. De måste alla slå den dömde på ryggen. Om det inte fanns 100 män måste man nöja sig med färre. Den som hade råd att betala böter kunde slippa löpa gatlopp.

Bödel hade staden inte. Bödeln var en man man fruktade. Han var så smutsig av sitt yrke att det t.ex. förekom att en dräng vägrade sova i samma säng som han. Drängen sov hellre på golvet. Bödel tillkallades vid behov från Vasa. Det fanns en galge i Gamlakarleby, på Galgbacken nära vattentornet, men den var sällan i användning.

* * *

1734 års lag blev den grund för rättens funktion som kom att till vissa delar gälla in till vår tid. Man hade dock redan från slutet av 1600-talet börjat hänvisa till lag i rättens utslag. Detta blev småningom regel i utslag. 1917 när Finland blev självständigt gällde 1734 års lag till betydande delar. Lagen var skriven för att läsas högt vid ting och överläggningar.

Straffen i lagen betraktades på den tiden som milda. Författaren Ivar Lo-Johansson kallar 1734 års lag ”de hängdas poesi”. Straff som höger hand avhuggen för stöld och annan stympning finns inte mera. Spö som straff gavs inte med ett utan med två sammanbundna spön för att kännas.

1734 fanns i missgärningsbalken ett kapitel 2 om trolldom och vidskepelse. Vi var inte långt från medeltiden. Lo-Johansson säger i förordet till en utgåva av den första versionen av lagen att också vår tid är hård och han frågar hur man om 200 år kommer att se på fängelserna.

* * *

Vid ett festligt tillfälle för Finlands riksdag i anledning av att Republiken firade sin etthundraårsdag i oktober 2017 sade riksdagens talman Maria Lohela vid ett möte i Stockholm med den svenska riksdagen att Finlands framgång inte varit möjlig om vi inte hade fått tron på demokratin, rättsstaten och jämlikheten i arv av Sverige. Hon sade i samma tal också att hon blivit mycket rörd när hon hört att Selma Lagerlöfs sista ord på sin dödsbädd våren 1940 när fredsavtalet ett par dagar tidigare hade slutits hade sagt: ”Hur har det gått för Finland?” Lohela höll sitt tal på finska och tolkades till svenska. Vid riksdagen i Finland går förhandlingarna på republikens båda språk.

Det svenska och nordiska arvet inom rätten är för oss idag en självklarhet men vi behöver inte resa långt i rum eller tid för att inse hur sällsynt vår syn på rättvisa och rätt är i världen.

Bibliografi

Humbley, J., Budin, G. & Christer, L. (eds.) (2018). *Language for special purposes. An international handbook*. DeGruyter.

Landqvist, H., Christer, L., Nordman, L., Nordman, M. & Kvist, M. (2016) *Juridik på svenska i Finland. Perspektiv på språk och rätt i Finland,* Scriptum.

Mattila, Heikki E. S.(2017) *Vertaileva oikeuslingvistiikka. Juridinen kielenkäyttö, lakimieslatina, kansainväliset oikeuskielet*. 2. Uud. Painos. ALMA TALENT.

Mattila, H. E. S., Pajula, S. & Piehl, A. (2010). *Oikeuskieli ja säädöstieto. Suomenkielinen lakikirja 250 vuotta. Rättsspråk och författningsinformration. Den finskspråkiga lagboken 250 år*. Suomalainen Lakimiesyhdistys.

Mickwitz, A. & Möllert, S. (1951). *Gamlakarleby stads historia*. *Del. I .Tidsskedet 1620-1713*. Gamlakarleby stads förlag.

Sveriges Rikes lag. (1981) *Gillad och antagen på Riksdagen år 1734. Faksimilutgåva med förord av Ivar-Lo Johansson*. Gidlunds.

Investigating test method effects in French L2 reading items for young learners

Peter Lenz, Katharina Karges and Malgorzata Barras
Institute of Multilingualism, University of Fribourg &
University of Teacher Education Fribourg

1. Introduction

1.1 Background

From 2014-16 the Competence Centre on Multilingualism carried out the 'Task Lab' study to have more firm ground for the upcoming item development for a country-wide computer-based survey (system monitoring) of sixth graders' receptive skills in their first foreign language learned at school.

The main emphasis of the 'Task Lab' project was on the exploration of specific design options for the French reading test. These options included, first, item type (test method) – short open answers (SA), multiple choice (MC) and matching (MTC); second, the language of the items – French, the target language, or German, the language of schooling. Although seemingly formal features, we suspected the choice of item type and language to have an influence on what is actually tested and, therefore, what the test scale stands for. The present paper focuses on the comparability of SA and MC items testing French reading skills in the CEFR A1 to A2+ range of levels.

1.2 Literature

There is a longstanding tradition of investigating test method effects in the field of education. Due to the wide use of MC items, they are often under scrutiny. Rodriguez (2003) performed a meta-analysis on the construct equivalence of MC and constructed-response[24] (CR) items. For this purpose, he formally summarised 56 correlations between MC and CR-based results. Almost 60 percent of these correlations stemmed from studies in language arts, the rest from various other fields. The main finding was that whenever item writers intended to tap the same construct using both item types, the test results on items of the two types were highly correlated. The average correlation

[24] 'Constructed response' stands in opposition to 'selected response'. Constructed responses may be short or extended. Multiple choice is one of several selected-response formats. In the following, when referring to studies, we use the terms for item formats that are used in these studies, e.g. 'open-ended items' for (a type of) constructed-response items.

turned out highest in studies that used stem-equivalent items, i.e. items using the same question, instruction or beginning of sentence to initiate the response process (MC or CR). The disattenuated correlation across studies amounted to 0.95 in this case.

In reading assessment, there is a tendency to use MC items to test lower-level skills and CR items to give test takers the opportunity to demonstrate higher-level reading skills such as global inferencing or reflecting on content. Obviously, in such circumstances construct equivalence between MC and CR items cannot be expected. Rauch and Hartig (2010) applies a two-dimensional latent regression model to investigate construct-differences between a general (L1) reading dimension, based on all MC and open-ended (OE) items, and a specific reading dimension, based on unaccounted variance from the OE items. The regression analyses showed several differential associations of social, cognitive and linguistic predictor variables with the two reading dimensions. However, it was impossible to attribute these findings with any certainty to item type because test method and construct(s) were confounded due to test design.

Ozuru and colleagues (Ozuru, Best, Bell, Witherspoon, & McNamara, 2007; Ozuru, Briner, Kurby, & McNamara, 2013) investigated construct equivalence of MC and OE items by using the same questions for both formats on the same tests. In all experiments described, the participants (U.S. college undergraduates) started by answering the OE items and then proceeded to the set of corresponding MC items. They were not allowed to go back and forth between OE and MC items. In the 2007 study (two experiments), half of the participants answered the OE and MC questions without having the opportunity to go back to the text passage, the other half could use the passage in the answering process. The results showed different test method effects depending on the availability of the passage. When the passage was unavailable, the effect size of the correlation between the scores based on the OE and the MC items respectively was large while it was only modest (and the correlation statistically nonsignificant) when the text was available during the response process. In the latter case, construct equivalence is doubtful. Ozuru et al. (2013) focuses on reading processes that might explain differential success on OE and MC items. While reading the text passage, the participants had to explain the meaning of some highlighted sentences in the text, thereby integrating information from different locations. After reading the passage, they first answered a series of OE items, then the corresponding series of MC items. When compared with the scores on both item types, the quality of the sentence explanations was moderately correlated with success on the OE items but not the MC items. The authors conclude that OE items measure more sensitively the quality of active generative processing during comprehension, while MC items tap in more passive recognition processes.

In L2-related research, Shohamy (1984) undertook an early systematic investigation of the effects of item design features on the measurement of the construct. She produced a total of eight English reading test versions by varying text prompt (two topics), test method (MC or OE items) and the language of the questions and options/answers (L1 Hebrew, the participants' stronger language, or L2 English). The

first part of the test was the same for all participants: Eight identical questions relating to the same eight passages served as a link between the test versions. Two main findings were that, on average, MC and L1 items were easier than OE and L2 items and that the effect of the harder conditions was stronger among less English-proficient students in the wide proficiency range represented in the sample.

1.3 Approach of the present study

Somewhat similarly to the Shohamy study, we also investigated test method effects by systematically varying the language of the items and the type of response. In addition, we collected information on precursor skills of reading as well as data from integrative tests that are usually strong correlates of reading comprehension.

The purpose of the present paper is to explore the equivalence of SA and MC items as test-methods for measuring the L2 French reading proficiency of young learners in the A1-A2+ level range. We refrain from the language-of-the-items issue and focus on item format effects using quantitative data from the main survey.

We investigate the following research questions:

a) Are there any systematic differences in the psychometric functioning of the SA and MC items used?

If there are differences –

b) how dramatic are they for the quality of a measurement instrument consisting of these item types?
c) in what way do the constructs represented by either of the two item types differ?

2. Method

2.1 Reading task development

We created the reading tasks for the study around 18 different text inputs. Twelve text inputs served as a basis for 36 (12 x 3) short-answer (SA) and 36 stem-equivalent multiple-choice (MC) items. The six remaining text inputs were used as a basis for 18 matching (MTC) items. Each of the SA, MC and MTC items came in two language versions, one with items in German[25] (the students' language of schooling), the other with items in French (the target language to be assessed). So, the complete test consisted of 144 SA or MC and 36 MTC (i.e. a total of 180) item versions[26].

[25] This means that all components of the items were in German, except for the text passages: in the case of SA, the question and the expected open answer; in the case of MC, the question and the three options; in the case of MTC, the question.
[26] As the matching items are quite different from the other items (no stem or content equivalence intended), we did not include them in this study.

Table 1. The task and item versions on the French reading test.

	short-answer	multiple-choice	matching
German items	French text inputs numbers 1-12 (12 x 3 items)	French text inputs numbers 1-12 (12 x 3 items)	French text inputs numbers 13-18 (6 x 3 items)
French items	French text inputs numbers 1-12 (12 x 3 items)	French text inputs numbers 1-12 (12 x 3 items)	French text inputs numbers 13-18 (6 x 3 items)

The tasks were designed as transfer-of-learning tasks for students who are all learning French in the same curricular region and with the same core of textbook materials. Task development followed a set of guidelines concerning types of reading (Urquhart & Weir, 1998) range of topics and the number of items per text input.

The writing of the SA and the MC items was marked by the decision to have all items in four versions by varying item language and item format. For example, the text input had to contain text references for the MC distracters and correct choices, even when the items were of the SA type. Conversely, all questions needed to be formulated precisely enough to narrow down the number of correct answers to one to have reliable short-answer items.

From a previous project, we had a corpus of all textbook materials available which the students had (at least potentially) worked with. Based on the corpus, we compiled a word frequency list, which served as a basis for component skills tests (e.g. vocabulary). We also used the corpus to check the familiarity of vocabulary items in text input and items.

2.2 Pre-piloting of reading tasks

The reading tasks were implemented in CBA ItemBuilder (DIPF & Nagarro IT Services, n.d.), a server-based test environment. Quality assurance was a major concern all along the test development and administration process. In a first phase, prototype reading tasks underwent usability testing to set the relevant screen design parameters and to improve functionality. After moderation by native speakers and experts, the reading tasks were pre-piloted by eight sixth grade classes[27]. We collected statistical routine information on the items and also a sample of the short answers we had to expect from the SA items (the "outcome space" according to Wilson, 2005) in order to prepare a coding key. In addition, we did one-on-one stimulated recall interviews (Gass & Mackey, 2000) with 34 students to collect evidence on the cognitive validity (Field, 2012) of our items.

[27] In Switzerland, sixth grade is the final grade of primary school. The great majority of students are between 11 and 12 years old. Average class size is slightly below 20. In primary school, the students of a class are normally taught together in all academic subjects. In the region we did our study, French teaching starts in third grade. From third to sixth grade, the average number of weekly French lessons amounts to 2.5.

2.3 Component skills assessments and integrative language tests

In addition to the reading tasks, we selected and developed a series of relatively short assessments of known correlates of reading comprehension, expecting that we could, among other things, use these additional measures to explore the construct or constructs embodied by the different types of items.

We settled for the following assessment instruments:

Table 2. Measurement instruments for component and reading task-related skills.

Test instrument		**Cognitive component(s) targeted**
1	Backward digit span task: repeat orally, in reverse order, a series of digits of increasing length	Working memory capacity (processing)
2	Read aloud French pseudowords	Phonemic awareness, French decoding/grapheme-to-phoneme conversion
3	Sight-word recognition	French sight-word reading; automatised receptive knowledge of whole written word forms
4	Yes/No Test	Breadth of French receptive vocabulary
5	Text segmentation (identifying word boundaries in text)	Receptive knowledge of French vocabulary and syntax; text segmentation accuracy
6	C-Test (integrative written gap-filling task)	French word/sentence/text comprehension in conditions of reduced redundancy; lexically and grammatically accurate word writing

Some brief comments on these assessment instruments:

1) Success on the backward digit span (BDS) task is a well-known predictor of success in reading, which, however, does not necessarily imply a substantive causality between working memory capacity and success in reading (Alderson et al., 2015). The BDS task is a simple and widely known working memory capacity (WMC) test that includes a secondary processing task (repeating the input backwards). Secondary processing also takes place when readers manipulate verbal information in their working memory. WMC accounts for a significant portion of variance in general intellectual ability (Conway et al., 2005). Our final BDS test included ten items, each two to six digits long, two of each length. Our students heard the ten series of digits in German on the computer headphones and repeated them orally.

2) The pseudoword reading aloud task assesses a learners' phonemic awareness and decoding skills in a language. Beginning readers rely a great deal on decoding. According to Geva and Siegel (2000) phonological and orthographic processing are

involved in decoding. The test uses pseudowords to make sure that grapheme-to-phoneme conversion actually needs to take place. We created the 20 items we needed to suit our student population with the help of a corpus-based web tool (New & Pallier, 2001).

3) The sight-word recognition task is a measure of sight-word reading, an advanced, automated form of word recognition that is crucial for fluent reading (cf. Alderson et al., 2015; Sabatini, Bruce, & Steinberg, 2013). We presented the students 20 French words (two to eight letters long). The words were visible on screen for just 80 milliseconds. Then the test takers spoke the words they had seen into a microphone.

4) The Yes/No Test (or Vocabulary Size Placement Test) (Meara & Buxton, 1987) is a well-known measure of receptive vocabulary breadth that is often used as a placement test. A Yes/No Test consists of real words and pseudowords. Test takers declare for every item they encounter whether they know it as a word of that language or not. The score on the pseudowords provides a false-alarm rate that can be used to correct the score on the existing words for guessing.

Our Yes/No Test consisted of 21 French words from the textbook corpus and 19 pseudo-French words that were generated in the same manner as the pseudowords for decoding.

5) The segmentation task is considered a combined (receptive) grammar-vocabulary task. In the DIALUKI study (Alderson et al., 2015), segmentation tasks proved to be strong predictors of reading proficiency. In a text segmentation task, test takers need to mark the word boundaries in one or more texts without blanks between the words.

6) The C-Test (Klein-Braley, 1985) is an integrative language test format whose strong association with language proficiency measures was established in many studies (e.g. Eckes & Grotjahn, 2006; Harsch & Hartig, 2016). A C-Test consists of a series of different texts (often four or five), in which, starting with the second word of the second sentence, the second half of every second (suitable) word is missing while the final sentence remains intact. Unlike the component skills tests, the C-Test involves written production of French, which is also the case for SA items. We used a C-Test from the Lingualevel collection (Lenz & Studer, 2007) with a total of 60 gaps.

We had the students of two classes do the six tasks described. Fourteen students from another class did the oral tasks (1-3), as well. In addition, they talked them through with a researcher in a one-on-one setting. The information gained in this manner helped to improve and customise the instruments.

2.4 Student questionnaire

The assessment instruments for the Task Lab study were accompanied by a short student questionnaire on social and language background, reading habits, language

learning motivation and perceived characteristics of the language teaching the students were experiencing. The items used in the questionnaire came from other questionnaires we had used in previous studies and were not pre-piloted again.

2.5 Piloting

For piloting the main survey, all data collection instruments (i.e. a brief questionnaire, the instruments presented in Table 2, and the reading tasks) were deployed on the CBA ItemBuilder system. This software allows access to customised test sets residing on a remote server by means of a current web browser. The goal for the reading test was to confront every student with a balanced sample of the existing task variants while never confronting the same student with the same task in two language or item format variants. Due to a time limit of 90 minutes for all written tasks, including the questionnaire, we confronted each student with a selection of 13 out of 18 available reading tasks (i.e. 39 items). A total of 24 different test sets was used. In these, the reading tasks appeared in different item format and language variants and positions. Overall, 119 sixth graders participated in these trial runs for the main survey.

2.6 The main survey

Overall, 609 sixth graders from 33 self-selected classes in 13 different schools located in German-speaking Switzerland were involved in the main data collection. All students were to do all tasks in the manner described for the piloting. Integral classes worked in the school's computer lab for 90 minutes, then went back to normal schoolwork. During the following lessons, small groups of students came to a separate room where they did the tasks with an oral component.

3. Results

We used the data obtained in the main study in various ways to identify differences, if they exist, with regard to a) the quality of the two item types as measurement instruments (3.1), and b) the constructs embodied by the two item types (3.2). While section 3.1 performs item analyses on the reading data, section 3.2 uses the results on the component and integrated measures tests as predictors of reading proficiency, measured separately by MC or SA items.

3.1 Format effects among the reading items

3.1.1 Data preparation and item selection

In a preliminary step, the answers to the SA items had to be coded. The provisional instrument from the pilot needed further refinement. Initial efforts to use partial-credit scoring were finally abandoned in favour of quasi-objectively applicable guidelines for dichotomous scoring. Interrater reliability was not evaluated statistically as all answers

were double-rated and all issues discussed in short intervals until mutual agreement between the two raters and a third person was reached.

Before investigating differences in the functioning of the SA and MC items for the present study, we first selected a set of quality items. For this purpose, all items were scaled[28] using the Rasch model and the 2PL (2-parameter logistic) IRT model[29]. There were between 83 and 154 (mean = 117.9) responses available per item variant. These relatively modest numbers are owed to the fact that each of the 609 students only solved a subset of 30 of the available 144 SA or MC item variants[30]. A total of 46 item variants was removed from the present analysis for various reasons (e.g. low discrimination, misfit). In order to diagnose misfit, mainly visual inspection of the empirical versus model item characteristic curve under the Rasch and the 2PL model was used, complemented by an inspection of the actual items and the answers provided. Whenever an item variant was excluded, its counterpart in terms of format (and language) was also excluded so that, now, for every MC item the corresponding SA item is also in the final set of items (and *vice versa*). The final set contains 98 item variants relating to 10 different passages; 588 students (290 females, 298 males) contributed usable responses. The Expected A Posteriori (EAP) reliabilities (Adams, 2005) amounted to 0.74 for the Rasch scale and 0.78 for the 2PL scale.

3.2 Analyses

A comparison of the difficulties of the items in both formats in the 2PL model reveals considerable differences.

Table 3. Mean difficulties of SA and MC items.

	short answer (SA)	multiple-choice (MC)
mean difficulty (logits)	1.349	-0.1256
SE (logits)	0.218	0.106

The difference is statistically significant on a paired t-test ($t = 7.67$, $df = 48$, $p < 0.001$), the standardised effect size ($d = 1.10$) large according to Cohen's rule-of-thumb interpretation. The findings for the item slopes (item discriminations in the 2PL model) are similarly clear:

[28] All statistical analyses were carried out using R software, for IRT the 'TAM' package (Kiefer, Robitzsch, & Wu, 2015).

[29] The Rasch model assumes that all items discriminate equally between weaker and stronger students. The 2PL model, however, estimates an individual discrimination parameter (the slope) for each item (Embretson & Reise, 2000). In order for the 2PL slope estimates to be stable over time, much larger numbers of test takers would be needed. However, generalisation of these parameters is not an issue here.

[30] Since four item variants were always based on the same question, the maximum workload would have been 36 items. Time limits made further reduction necessary.

Table 4. Mean slopes of SA and MC items under the 2PL model.

	short answer (SA)	multiple-choice (MC)
mean 2PL slope	1.535	0.657
SE	0.091	0.041

A difference of 0.878 is statistically significant on a paired t-test ($t = 9.26$, $df = 48$, $p < 0.001$), and the effect size ($d = 1.32$), again, is large. A difference in slope (i.e. discrimination) between SA and MC items is not unexpected considering what it generally takes to answer items of either type. In the case of SA items, it is not enough to understand and answer a question – the answer also needs to be formulated and written down (in German or French, depending on the item variant), otherwise comprehension remains unnoted. In the case of the purely receptive MC items, better students have fewer opportunities to prove that they actually are better. The theoretical 33% chance of guessing the right answer mitigates the power of an item to discriminate between weaker and stronger test takers and so does the fact that comprehension can be documented by simply ticking a box.

If person measures are produced based on a 2PL model, the slope or discrimination parameter is used to weight the scores on the individual items. So, getting an item right or wrong, counts more if the slope of an item is steeper. In the present case, the average SA item would contribute more than twice as much as the average MC item to the weighted person scores.

The frequently used Rasch model assumes equal slopes and therefore weights every item equally. Consequently, the raw score (number of correctly solved items) is considered a sufficient statistic. Test takers who have a higher total score on the same test, no matter which of the items they solved correctly, have higher ability according to the model. In addition to this principle of sufficiency of the raw score, Rasch (Rasch, 1977) postulates the related principle of specific objectivity as a fundamental property of the Rasch model: Any sub-sample of items from this test would classify any sub-group of test takers in the same order. From a Rasch measurement perspective, our findings regarding the two item types indicate a (undesirable) case of differential item group functioning. In practice, differential item or item group functioning is commonly observed due to person or item groups that have something in common others do not have. Profile Analysis (Verhelst, 2011; Yildirim, Yildirim, & Verhelst, 2014) provides the statistical means to evaluate the strength and significance of such effects.

Our statistic of interest was the mean deviation profile for several ability groups that we formed along the common Rasch scale constructed from our SA and MC reading items. An individual deviation profile is calculated as follows: The expected score, based on the Rasch model, on the completed items of each item group (SA or MC) is subtracted from the observed score on the items of each group. The differences on all item groups (here two) form the deviation profile. The mean deviation profile is an aggregation of the individual deviation profiles of the test takers per ability group. For our analysis we defined three ability groups based on the Rasch scale: a middle group including person scores +/- 0.5 SDs around the mean, and the two groups left

and right of this band. The actual group sizes were 161 (weakest group), 257 and 155 students (28%, 45%, 27%). 15 students were excluded from the analysis because they had either extreme scores or no data on one of the two item types.

The resulting mean deviation profiles show highly significant deviations from the Rasch model-based score predictions for the lowest and the highest-scoring groups (cf. Table 5).

The results for the three ability groups show how much the average of the observed scores differs from the average of the expected scores in each item group. The results for the two item types add up to zero in each ability group. The least-ability group scored significantly higher than predicted by the model on the MC items while the highest-ability group scored significantly higher than expected on the SA items. Evidently, the contrast between these two ability groups on the two item types is even larger than the difference between the observed and the expected mean for each group.

Table 5. Mean deviation profile for three ability and two item groups.

Ability group	SA items	MC items	SE	z	p
lowest	-0.394	0.394	0.062	-6.352	< 0.001
middle	-0.004	0.004	0.056	-0.064	0.475
highest	0.376	-0.376	0.073	5.159	< 0.001
lowest - highest	-0.770	0.770	0.096	-8.056	< 0.001

The above findings show that the assumption of specific objectivity is not appropriate for our set of items, nor is the test score a sufficient indicator for a person's ability. Depending on the sub-sample of items (esp. types of items) they are confronted with, the Rasch model may classify test takers in different ability groups either too low or too high on the latent ability scale.

3.2.1 Exploring construct-equivalence of SA and MC items

In order to explore potential systematic differences in the demands the items in both formats make, we used the results of the component skills and integrative tests to find associations between the constructs they embody and the constructs underlying the SA and MC-based reading tests (similarly: Rauch & Hartig, 2010). For this purpose, we first scaled the reading data using yet another IRT model, and prepared the scores from the different component skills and integrative measurements for further analysis. Then, we combined them with other (i.e. structural and questionnaire) variables in a single dataset, and performed multiple imputation on this dataset. Multiple imputation produced a series of complete datasets that could easily be used in multiple regression analysis to explore associations between the component skills or integrative tests (independent variables) and reading comprehension through SA or MC items (dependent variables).

3.3 Data preparation

In order to suit the purpose of this part of the study, the SA-based and the MC-based reading scores were additionally scaled using two-dimensional Rasch analysis (Reckase, 2009). Dimension 1 (EAP reliability 0.74) was based on the SA items, dimension 2 (EAP reliability 0.69) on the MC items[31]. The latent correlation between both dimensions amounted to 0.91, suggesting closely related constructs. WLEs (Warm's weighted likelihood estimates, Warm, 1989) were output as person estimates for subsequent use.

For the backward digit span task, we defined the score as the length of the longest string of numbers the students correctly repeated backwards. The maximum string-length metric (ML) is one of two metrics Woods et al. (2011) recommend based on their comparative study.

Coming up with a coding scheme for the decoding task proved challenging. We finally settled on 37 different syllables as our items. A single rater coded them once. The Rasch scale produced had an EAP reliability of 0.86. Again, WLEs were estimated as person measures.

From the 20 sight-word recognition items we selected 14 for coding, the six remaining being too easy. We managed to apply partial credit scoring quasi-objectively. The items were Rasch scaled (EAP reliability = 0.78), and, again, WLE person estimates were produced.

When Yes/No Test scores are used in practice, the number of 'yes' on the existing-word items is usually corrected by the number of 'yes' on pseudoword items (false-alarm score or rate) (cf. Huibregtse, Admiraal, & Meara, 2002). Without a correction, a test taker could attain the maximum score by simply choosing 'yes' for every item. Following a recommendation by Harsch & Hartig (2016), we constructed separate measures for the 21 words and the 19 pseudowords using a two-dimensional Rasch model (EAP reliabilities: words 0.77; pseudowords 0.70). For subsequent analyses, two WLE scales were output.

The text segmentation scale was produced by counting the correct segmentations in each text (after some exclusions). We standardised both score scales, added the resulting values and standardised the sums again to get the final standardised score.

In the case of the C-Test, each of the three texts was treated like a polytomous item with up to 17 categories after collapsing a few categories with too sparse data. The items were Rasch scaled (EAP reliability = 0.70), and WLEs were produced as person measures.

To prepare for data imputation, all test scores and scales were merged into a single dataset and complemented with indicator variables (e.g. students' class membership) and variables from the student questionnaire covering topics such as reading habits and motivations for learning French. On this dataset, we performed

[31] Thanks to this split, the problems with the Rasch scale detected in the previous section are not an issue here.

multiple imputation using the R package 'mice' (multivariate imputation by chained equations, Buuren & Groothuis-Oudshoorn, 2011). Our data imputation had two objectives: first, replacing missing data with plausible data, and second, accurately representing person measures containing measurement error (here the WLEs). For the purpose of the present study, 240 imputed datasets were produced, from which we used 40, i.e. every sixth set, in the data analysis[32]. All statistical analyses based on imputed datasets need to be carried out 40 times[33] independently. The results are combined according to Rubin's rules (Rubin, 1987).

3.4 Multiple regression analyses

The intercorrelations of the cognitive (backward digit span) and the various language-related predictor and criterion variables afford an overview of existing associations between variables.

Table 6. Correlations between cognitive and ling. variables (mean correlations from 40 imp. sets).

	De--coding	S-w recog.	Y/N words	Y/N pseud.	Y/N diff.	Text segm.	C-Test	Read. SA	Read. MC
Backward digit span (z)	0.18	0.28	0.05	-0.11	0.18	0.13	0.13	0.23	0.18
Decoding (z)		**0.77**	0.44	-0.23	**0.72**	0.66	0.67	0.51	0.48
Sight-word recognition (z)			0.42	-0.26	**0.74**	0.61	0.66	0.56	0.52
Y/N Test, words (z)				0.58	0.46	0.39	0.41	0.43	0.40
Y/N Test, pseudowords (z)					-0.46	-0.29	-0.30	-0.14	-0.25
Y/N Test, difference (z)						**0.75**	**0.78**	0.62	**0.70**
Text segmentation (z)							**0.84**	0.56	0.52
C-Test (z)								0.58	0.51
Reading SA items									0.63

The Pearson product-moment correlations presented in Table 6 are averages from the 40 imputation sets. The highest correlations (> 0.7) are highlighted in bold type. In most of these high correlations, the ad-hoc variable 'Y/N Test, difference (z)' is involved. This is the standardised difference between the standardised 'words' score and the standardised 'pseudowords' score of the Yes/No Test. Neither of these two shows strong associations with any of the other variables, but the difference does. This difference also shows the strongest association with either of the item type-specific reading scores. With the MC reading subtest it shares nearly 50% of the variance

[32] In each of these datasets, the former WLE measures differ slightly as they were drawn from the error distribution of the original WLE measures during imputation.
[33] We chose to work with such a high number of imputed datasets because of the partly only moderate scale reliabilities. More datasets can better reflect a wider error distribution (uncertainty) of the person measures.

(squared correlation $R^2 = 0.49$). Another noteworthy observation is the fact that the short and simple phonemic awareness/decoding test and the sight-word recognition test show similarly strong associations with both reading subscores as the more integrative text segmentation task and the C-Test.

In order to explore to what extent each of the component skills and integrative tests share variance with the SA-based and the MC-based reading test, we used stepwise multiple regression (cf. Sabatini, O'Reilly, Halderman, & Bruce, 2014). Building on a background model (containing some social and attitudinal variables), we first added the scores from the cognitive backward digit span task, then the fundamental, language and reading-related decoding and sight-word recognition tasks, and so on, until we finally arrived at the measure from the integrative C-Test. With every additional variable, we recorded the variance shared (R^2) between the updated model and the reading measures as well as the AIC. Since we chose a linear mixed-effects model (LMM)[34], we actually used a (marginal[35]) pseudo-R^2 based on Nakagawa & Schielzeth (2013), implemented in the R package 'piecewiseSEM' (Lefcheck, 2016). The numbers reported in Table 7 represent the mean of the results obtained from our 40 datasets. The AIC (Akaike Information Criterion) is an indicator of the fit of the model to the data (lower numbers mean better fit). The AIC statistic also penalizes higher numbers of predictor variables, i.e. it favours leaner models to some degree.

The predictors marked with an asterisk (*) in Table 7 all significantly improved the multiple regression models for both reading scales when they were first introduced in the given order. With the introduction of (correlated) predictors that capture similar but more comprehensive reading-related skills, the earlier predictors essentially lost this function and turned insignificant except for the best predictors (for details see Table 11).

[34] The LMMs were estimated with the 'lmer' function from the 'lme4' R package (Bates, Mächler, Bolker, & Walker, 2015). The pooling was done with the 'pool' function from the 'mice' package (Buuren & Groothuis-Oudshoorn, 2011).

[35] The marginal R^2 takes into account the variance shared between the fixed effects (background and predictor variables) and the reading scores but not the variance explained by the random effect (school classes).

Table 7. Results of stepwise, hierarchical multiple regression for SA and MC reading items.

	SA reading items				MC reading items			
	R^2	R^2 Change	AIC	AIC change	R^2	R^2 Change	AIC	AIC change
Background variables	0.157	-	6960.6	-	0.107	-	6864.4	-
Backward digit span (z)*	0.196	0.039	6933.8	-26.8	0.13	0.023	6834.2	-30.2
Decoding (z)*	0.335	**0.139**	6813.4	-120.4	0.262	**0.132**	6751.3	-82.9
Sight-word recognition (z)*	0.389	0.054	6780.0	-33.4	0.309	0.047	6704.0	-47.3
Y/N Test, words (z)*	0.417	0.028	6761.0	-19.0	0.337	0.028	6701.4	-2.6
Y/N Test, pseudowords (z)*	0.486	**0.069**	6734.0	-27.0	0.574	**0.237**	6530.2	-171.2
Text segmentation (z)	0.504	0.018	6690.7	-43.3	0.577	0.003	6526.8	-3.4
C-Test (z)	0.516	0.012	6679.5	-11.3	0.584	0.007	6528.5	1.7

Overall, we find that our predictors share more variance (R^2) with the MC reading measure than with the SA reading measure (58.4% vs. 51.6%). This is mainly due to the Y/N Test. When it is added to the model, it contributes an additional 26.5% of shared variances in the case of the MC-based test but 'only' an additional 9.7% in the case of the SA-based test. Text segmentation and the C-Test seem irrelevant as further predictors of the MC test result but keep improving the fit of the model (AIC) that predicts the SA reading score, even though the additional 3% of shared variance seems modest. A closer look (Table 8) at the two scores derived from the Y/N Test reveals that neither one of them is an extraordinary predictor by itself (as the moderate correlations already suggested) but that together they make a great and differential impact on our models.

Table 8. The two Y/N Test dimensions as predictors in reverse order (cf. Table 7).

	SA items				MC items			
	R^2	R^2 Change	AIC	AIC change	R^2	R^2 Change	AIC	AIC change
Sight-word recognition (z)	0.389	0.054	6780.0	-33.4	0.309	0.047	6704.0	-47.3
Y/N Test, pseudowords (z)	0.428	0.039	6780.9	0.9	0.332	0.023	6686.0	-18.0
Y/N Test, words (z)	0.486	0.058	6734.0	-46.9	0.574	**0.242**	6530.2	**-155.8**

If the Y/N Test is excluded from the set of predictor variables, a big difference between the models for the SA and the MC items becomes visible (Table 9, Table 10).

Table 9. Hierarchical multiple regression without the Y/N Test; text segmentation added, then C-Test.

	SA items				MC items			
	R^2	R^2 Change	AIC	AIC change	R^2	R^2 Change	AIC	AIC change
Sight-word recognition (z)	0.389	0.054	6780.0	-33.4	0.309	0.047	6704.0	-47.3
Text segmentation (z)	0.448	0.059	6706.0	-74.0	0.361	0.052	6645.5	-58.5
C-Test (z)	0.474	0.026	6691.9	-14.1	0.371	0.010	6643.2	-2.3

Table 10. Hierarchical multiple regression without the Y/N Test; C-Test added, then text segmentation.

	SA items				MC items			
	R^2	R^2 Change	AIC	AIC change	R^2	R^2 Change	AIC	AIC change
Sight-word recognition (z)	0.389	0.054	6780.0	-33.4	0.309	0.047	6704.0	-47.3
C-Test (z)	0.465	0.076	6703.3	-76.7	0.355	0.046	6658.8	-45.2
Text segmentation (z)	0.474	0.009	6691.9	-11.4	0.371	0.016	6643.2	-15.6

Concerning the SA-based reading test, the highly correlated ($r = 0.84$) text segmentation and C-Test measures together add an amount of shared variance to the model that is comparable to the contribution the Y/N test makes. With regard to the MC-based test, however, their explanatory power remains modest. Text segmentation and the C-Test together add a mere 6.2% of shared variance while the two Y/N Test measures add 26.5%. In the SA model, text segmentation appears almost redundant as a predictor when added second (Table10), in the MC model the same is true for the C-Test (Table 9). In the regression model for SA-based reading that comprises text segmentation and the C-Test but excludes the Y/N test (output not shown here), the C-Test measure improves the model significantly at the 95% confidence level ($t = 2.30$, $p = 0.023$) while the text segmentation score only just reaches borderline significance ($t = 1.85$, $p = 0.065$). In the corresponding model for MC-based reading, it is the reverse situation, just clearer: text segmentation is a significant predictor ($t = 2.25$, $p = 0.026$) while the C-Test is not ($t = 1.16$, $p = 0.248$).

In order to evaluate the differential effects the predictors have on SA and MC-based reading by means of inferential statistics, we estimated a joint LMM model in which the SA reading score and the MC reading score are implemented as repeated measures while all predictors interact with item type.

The left and right-hand panels in Table 11 provide extracts from the model output (i.e. fixed effects parameters of interest) of two equivalent multiple regression models. The upper left panel shows the cognitive and language-related fixed effects predictors for the SA reading score. The lower left panel adds the interaction effects

representing the 'corrections' that have to be made to the predictors for SA-based reading in order to optimally predict MC-based reading. The p-values in bold type in the left panel indicate that the backward digit span score and both Y/N Test scores are significant predictors of SA-based reading. In addition, the significant interaction effects for the Y/N Test scores statistically endorse the observation that the Y/N measures are differentially associated with the two reading scores. Such a differential effect is not confirmed for the C-Test as a predictor. The right-hand panel displays the same results as the left-hand panel but takes the main effects for the prediction of MC-based reading as a point of departure. It shows that the Y/N Test measures are the only statistically significant predictors for MC-based reading in our set.

The effect size of the predictors can be directly inferred from these tables because we entered the reading measures on scales with a standard deviation (SD) of 100 while all predictors were coerced to a standardised scale (mean = 0, SD = 1). So, for example, a score of 0.5 instead of -0.5 on the 'words' dimension of the Y/N Test predicts an MC-based reading measure that is more than 1 SD (115.88 units) higher.

Table 11. Cognitive and language-related predictors of reading (SA-based vs. MC-based).

	SA reading measure					MC reading measure				
	coeff.	SE	t	df	p	coeff.	SE	t	df	p
Main effects (extract)										
Backward digit span (z)	8.62	4.22	2.04	113.1	**0.042**	4.42	5.56	0.79	60.0	0.428
Decoding (z)	-10.26	14.38	-0.71	37.2	0.476	-13.63	18.45	-0.74	29.6	0.461
Sight-word recognition (z)	4.32	17.65	0.24	34.2	0.807	-9.58	21.13	-0.45	29.0	0.651
Y/N Test, words (z)	63.08	28.58	2.21	29.0	**0.028**	115.88	33.80	3.43	24.8	**0.001**
Y/N Test, pseudowords (z)	-46.84	27.68	-1.69	27.3	**0.092**	-103.23	31.19	-3.31	24.4	**0.001**
Text segmentation (z)	11.55	12.35	0.93	52.2	0.351	4.41	13.43	0.33	47.0	0.743
C-Test (z)	14.12	15.81	0.89	44.9	0.373	-11.95	16.73	-0.71	42.3	0.476
Interactions: item type x predictors (extract from output)										
	'Correction' for MC measures					'Correction' for SA measures				
Backward digit span (z)	-4.20	5.95	-0.71	75.56	0.48	4.20	5.95	0.71	75.56	0.48
Y/N Test, words (z)	52.80	25.98	2.03	33.8	**0.043**	-52.80	25.98	-2.03	33.8	**0.044**
Y/N Test, pseudowords (z)	-56.39	25.04	-2.25	31.8	**0.025**	56.39	25.04	2.25	31.8	**0.026**
Text segmentation (z)	-7.14	13.31	-0.54	56.2	0.592	7.14	13.31	0.54	56.2	0.593
C-Test (z)	-26.07	16.70	-1.56	46.7	0.120	26.07	16.70	1.56	46.7	0.121

4. Discussion

Psychometric item analysis of our stem-equivalent MC and SA reading items revealed large and significant differences in the functioning of MC and SA items with regard to difficulty and discrimination. Our study confirms Shohamy's (1984) findings in a similarly designed study that MC items are easier than SA items. We assume that the relatively high probability of 0.33 of guessing the correct MC option as well as the fact that the answering process involves fewer (or no) productive elements can serve as a general explanation.

The average discrimination of the SA items is more than double the discrimination of the MC items. Generally, if an item has low discrimination in relation to a scale, it has a weak relationship with the specific dimension defined by all the other items (Wilson & Hoskens, 2005). In our case, the difference is particularly remarkable because the complete test consists of an equal number of these two types of item so that, in principle, both could equally contribute to the common dimension (construct). Apparently, the contribution the MC items make, is diluted while the opposite is true for the SA items. It seems likely that a range of different test taking strategies can be applied in the case of the MC items, which tap less intensively and less uniformly into language-related resources than successful test-taking strategies for SA items do as they involve active understanding (no answers suggested) as well as active (productive) answering, both involving language resources.

The latent ('error free') correlation (Wu, Adams, Wilson, & Haldane, 2007) of 0.91 between the MC and SA reading dimensions estimated by the two-dimensional Rasch model is roughly equivalent to the average 0.95 disattenuated correlation of stem-equivalent SA and MC items in Rodriguez' (2003) meta-study. The magnitude of this correlation suggests that a test consisting of both items types is essentially uni-dimensional, i.e. it is appropriate to measure a common construct. We could not necessarily expect high correlation because in our test, text and items were concurrently present, a constellation that did not result in a significant correlation between SA-based and MC-based scores in the study by Ozuru et al. (2007).

However, Profile Analysis reveals that care needs to be taken when MC and SA items are used on the same Rasch scale because they are the source of significant non-uniform differential item group functioning[36]. Concretely, weaker students get a relative advantage from MC items while SA items benefit stronger students (and vice versa) when all items have the same weight, which is the case in Rasch measurement. This issue can be resolved most notably by applying the 2-parameter logistic IRT model instead of the Rasch model. The 2PL model uses item-specific weights and thus takes into account the strength of the relationship an item has with the latent measurement dimension. In the present case, applying the 2PL model instead of the Rasch model increases the variance of the person (WLE) scale by roughly 20%.

[36] The group of SA items and the group of MC items distort the measures to changing degrees (i.e. non-uniformly) along the ability scale.

Our attempt to shed light on differences between the SA and MC reading constructs by means of regression modeling has been somewhat successful. Vocabulary breadth is the best predictor of reading success on both reading scales, but it is even a significantly better predictor with regard to MC-based reading (Table 11). In the complete model for MC-based reading, no other predictor reaches statistical significance. In the model for SA-based reading, however, the backward digit span score also reaches significance. Also, when vocabulary breadth is replaced by the C-Test score, the total variance which the model shares with SA-based reading is only 2.1 percentage points lower (46.5% vs. 48.6%).

The explanatory power of a vocabulary test as such serves as no surprise because in the A1-A2 range of levels reading is usually found to be more of a language than a reading problem (Alderson, Nieminen, & Huhta, 2016; Alderson & Urquhart, 1984). It is tempting to speculate about commonalities between MC-based reading and the Y/N Test. Being successful on our Y/N Test implies the ability to recognise words already encountered before with some certainty. Existing words should not be missed while pseudowords should be discarded. Success on MC items similarly depends on an interplay between selection and deselection based on recognition. SA-based reading on the other hand comprises a productive element – formulating and writing an answer, be it in German or French. This may explain the observed association with the C-Test score. The fact that working memory capacity is a significant predictor of SA-based reading recalls Ozuru et al.'s (2013) finding that success on OE items depends on active generative processing of the input text.

4.1 Limitations and outlook

Our study concerns quite a specific population: German-speaking sixth graders learning basic French in a school context. Research involving learners with more advanced literacy skills, more elaborate test-taking strategies and higher L2 language ability might come to partly different conclusions. Also, the kind of statistical evidence we collected should be complemented by data from introspective research methods and particularly eye-tracking (Brunfaut, 2016; Brunfaut & McCray, 2015) to attain a richer understanding of test takers' actual reading and problem-solving processes when answering SA and MC items.

The set of predictors of reading ability is another point to improve. In order to pinpoint differences on item types and facilitate interpretation, there should be more measures capturing specific component or precursor skills (cf. Alderson et al., 2015).

In addition, the mostly statistical approach we chose takes little notice of individual item characteristics. It would be beneficial if the interplay of text and item characteristics was generally better understood. Item difficulty or discrimination modelling that goes beyond simple test method factors, could greatly enhance the knowledge base item developers can draw on. With respect to our data, so-called retrofitting of task factors will be a logical next step.

5. Conclusion

In our study, we used stem-equivalent SA and MC reading items to explore the equivalence of both item types with respect to scale quality and construct representation in a French-as-an-L2 context with young learners at an elementary ability level. We could show that SA items are, on average, better representatives of the measurement scale embodied by an equal number of SA and MC items. The presence of significant differential item group functioning confirmed through Profile Analysis suggests that simple Rasch scaling is problematic in the presence of SA and MC items because all items are weighted equally.

A latent correlation larger than 0.9 between SA-based and MC-based reading indicate that, overall, both tests methods measure the same construct. As expected for L2 readers at low language ability levels, receptive vocabulary knowledge is the best predictor of reading success, especially when reading is measured through MC items. The fact that working memory capacity is the only other concurrently significant predictor of SA-based reading, may indicate that more active generative processing is involved in answering short-answer items.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*(2–3), 162–172.

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L. & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. New York: Routledge.

Alderson, J. C., Huhta, A. & Nieminen, L. (2016). Characteristics of weak and strong readers in a foreign language. *The Modern Language Journal*, *100*(4), 853–879.

Alderson, J. C. & Urquhart, A. H. (1984). Reading in a foreign language: A reading problem or a language problem? In *Reading in a foreign language* (pp. 1–24). London: Longman.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Brunfaut, T. (2016). *Looking ino reading II: A follow-up study on test-takers' cognitive processes while completing APTIS B1 reading tasks*. British Council. Retrieved from https://www.britishcouncil.org/sites/default/files/brunfaut_final_with_hyperlinks_3.pdf

Brunfaut, T. & McCray, G. (2015). *Looking into test-takers' cognitive processes while completing reading tasks* (ARAGs Research Reports Online No. AR/2015/001). British Council. Retrieved from https://www.britishcouncil.org/sites/default/files/brunfaut-and-mccray-report_final.pdf

Buuren, S. van & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 67.

Conway, A. A., Kane, M., Bunting, M., Hambrick, D. Z., Wilhelm, O. & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

DIPF, & Nagarro IT Services. (n.d.). *CBA ItemBuilder*. Frankfurt (Main). Retrieved from https://tba.dipf.de/de/infrastruktur/softwareentwicklung/cba-item-builder/cba-itembuilder

Eckes, T. & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*(3), 290–325.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J: L. Erlbaum Associates.

Field, J. (2012). A cognitive validation of the lecture-listening component of the IELTS listening paper. In L. Taylor & C. J. Weir (Eds.), *IELTS collected papers 2: Research in reading and listening assessment* (pp. 17–65). New York, Cambridge: Cambridge University Press.

Gass, S. M. & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Routledge.

Geva, E. & Siegel, L. S. (2000). Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages. *Reading and Writing*, *12*(1), 1–30.

Harsch, C. & Hartig, J. (2016). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing*, *33*(4), 555–575.

Huibregtse, I., Admiraal, W. & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, *19*(3), 227–245.

Kiefer, T., Robitzsch, A. & Wu, M. (2015). *TAM: Test Analysis Modules*. Retrieved from http://CRAN.R-project.org/package=TAM

Klein-Braley, C. (1985). A cloze-up on the C-Test: a study in the construct validation of authentic tests. *Language Testing*, *2*(1), 76–104.

Lefcheck, J. S. (2016). piecewiseSEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, *7*(5), 573–579.

Lenz, P. & Studer, T. (2007). *lingualevel: Französisch und Englisch. Instrumente zur Evaluation von Fremdsprachenkompetenzen* (1.). Schulverlag.

Meara, P. & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142–154.

Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

New, B. & Pallier, C. (2001). Lexique Toolbox - Nonmots, pseudomots, voisins - des outils pour la psycholinguistique. Retrieved January 4, 2016, from http://www.lexique.org/toolbox/toolbox.pub/

Ozuru, Y., Best, R., Bell, C., Witherspoon, A. & McNamara, D. S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction*, *25*(4), 399–438.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *67*(3), 215-227.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In *The Danish Yearbook of Philosophy* (Vol. 14, pp. 58–93). Copenhagen: Munksgaard. Retrieved from https://www.rasch.org/memo18.htm

Rauch, D. & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, (4), 354–379.

Reckase, M. (2009). *Multidimensional item response theory*. New York; London: Springer,.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: a random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163–184.

Sabatini, J. P., Bruce, K. & Steinberg, J. (2013). SARA reading components tests, RISE form: test design and technical adequacy. *ETS Research Report Series*, *2013*(1), i–25.

Sabatini, J. P., O'Reilly, T., Halderman, L. K. & Bruce, K. (2014). Integrating scenario-based and component reading skill measures to understand the reading behavior of struggling readers. *Learning Disabilities Research & Practice*, *29*(1), 36–43.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, *1*(2), 147–170.

Urquhart, A. H. & Weir, C. J. (1998). *Reading in a second language: process, product, and practice*. London, New York: Longman.

Verhelst, N. D. (2011). Profile Analysis: a closer look at the PISA 2000 reading data. *Scandinavian Journal of Educational Research*, 1–18.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Mahwah N.J.: Lawrence Erlbaum Associates.

Wilson, M. & Hoskens. (2005). Multidimensional item responses: Multimethod-multitrait perspectives. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: a book of exemplars : papers in honour of John P. Keeves* (pp. 287–307). Dordrecht: Springer.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. (2007). *ACER ConQuest Version 2*. Melbourne: ACER Press.

Yildirim, H. H., Yildirim, S. & Verhelst, N. (2014). Profile Analysis as a generalized differential item functioning analysis method. *Education and Science*, *39*(214), 49–64.

What counts as language proficiency for UK citizenship: The B1 Benchmark?

Constant Leung
King's College London

Jo Lewkowicz
University of Warsaw

1. Introduction and Contextualization

In the past decade we have seen a resurgence of the prominence of the notion of 'country' as a political unit that controls and regulates people's cross-border movements and settlements. Recent events in Europe provided a vivid demonstration of the extraordinary role played by individual countries, in the guise of nation states, in facilitating and/or hindering the movement of people. For instance, in 2016 it was reported in the press that Germany took in some 900,000 refugees from war-torn Iraq, Syria, and other turmoiled places (Washington Post, 30/09/2016 https://www.washingtonpost.com/news/worldviews/wp/2016/09/30/germany-said-it-took-in-more-than-1-million-refugees-last-year-but-it-didnt/?utm_term=.3bed4c4a560a).There were nightly television news clips showing groups of bedraggled people travelling through Europe making their way from Italy and Greece in the south, northwards to Germany, Sweden and other northern countries on foot or in buses and trains. Along the way some countries refused entry, others created special through-pass corridors to ensure that the unwanted refugees would not stay in those 'pass-through' countries. If nothing else, this dramatic episode of mass movements of people in our time vividly illustrates the power of the sovereign country in determining who has the right to come and stay, and who might not. The idea of a country in these migration-related matters is something of an abstraction. It is, of course, the governments of the various European countries that are making the gate-keeping decisions; it is those same governments whose political authority derives from the claim that they act on behalf of their citizens.

The focus of this chapter is on the ways in which such gate-keeping decisions regarding the issuance of permanent citizenship are made in relation to language in the UK, which has a complex immigration control policy. Gaining permanent residency is connected to (the subsequent option of) naturalization as a British citizen, which is closely linked to the government concerns for social integration. In this discussion we will side-step the legal requirements and administrative processes necessary for different groups or categories of individuals to gain residency, as they are not central

to our purpose. Instead, we will briefly describe the social context in which, as part of the application for permanent residency, the passing of an English language test is a pre-requisite.

Since the 1950s, (the scale of) immigration into the UK has been a 'sensitive' political and social issue in that it has always attracted strongly expressed opposition in certain quarters in the population. The arrival of British subjects from the Commonwealth countries in the 1950s and 1960s, the East African British Asians in the 1970s, and the EU citizens in the 1990s and 2000s all triggered public support and criticism at the same time. The perceived lack of social integration, partly attributed to the lack of English language fluency, and the dilution of available social services and housing caused by the incoming groups are usually presented in the public media as the main reasons for saying 'too many immigrants' and 'no more' (see Blackledge, 2009; Webster, 2018 for historical background). More recently, the violent actions committed in the name of some Islamic Jihadist groups in different parts of the world, including the UK, have played into the anti-immigration narrative. The already established narrative of 'lack of integration' is further strengthened by the additional anti-immigrant rhetoric of 'lack of security because of the strangers in our midst'. The obverse side of this narrative is that security in society requires all citizens to share similar social values and practices within a common culture, and speaking English is a key ingredient of this sharing of a common culture. In sum, the granting of permanent residency means, inter alia, the applicant is required to demonstrate an ability to speak English. The putative causal relationship between sharing culture and language has been questioned (Blackledge, 2009; Kunnan, 2010; McNamara, 2005), and the linear correlation between language proficiency, peace, and safety in society is, of course, not as straightforward as the assumptions underpinning the current assessment requirements (Van Avermaet and Gysen, 2009; Gostjev and Nielsen, 2016). Nevertheless, this is the discourse adopted by the UK Government and enacted through the authority of the Nationality, Immigration and Asylum Act 2002. (We will return to this issue in the final discussion). It should be noted though that the UK Government is certainly not alone in adopting this kind of policy disposition; many other governments have adopted similar approaches (Bruzos et al, 2017; Extra et al, 2009).

In this chapter the focus of our attention is on the English language requirement itself. More specifically, we are interested in the kinds of language use being sampled in the test. At the present time there are two officially approved tests of spoken English for the granting of permanent residence, also known as Secure English Language Tests: IELTS B1 level Life Skills, and Trinity College London Graded Examinations of Spoken English (GESE) Grade 5 (https://www.gov.uk/guidance/immigration-rules/immigration-rules-appendix-o-approved-english-language-tests). Both of these tests are said to be referenced to CEFR B1 level. In the next section we provide a brief characterisation of B1-ness as embedded in the CEFR. We will then describe the two secure tests, paving the way to analysing the interactional moves performed by the test participants. In the final part of the paper, we will comment on the adequacy of the tests in terms of construct validity, fitness for purpose and policy viability.

2. CEFR B1 Spoken English

What does CEFR B1 level proficiency comprise in terms of language knowledge and use? In the absence of a summary definition or description of each of the levels within the CEFR, it would seem that the first step in composing a holistic picture is by assembling the relevant B1 descriptors in the first three scales in the CEFR (Council of Europe, 2001:24-29): Global Scale, Self-assessment Grid, and Qualitative Aspects of Spoken Language use, as these set out the common reference levels for all the supplementary scales within the CEFR framework (including the additional descriptors in the Companion Volume, CEFR, 2018).

For spoken language use, CEFR B1 level is manifested in the descriptors set out in Table 1. An analysis of the three 'master scales' set out in Table 1 would suggest there are a number of key components that characterize and underpin the CEFR B1 speaking construct, allowing for the following characterisation of 'B1ness':

The CEFR B1 descriptors depict an adult who is at the same time learner, speaker and hearer of an additional language. They have the ability to engage and communicate through speaking and listening, though they remain an outsider to the local speech community of the target language. They can communicate on familiar topics and those of direct everyday interest to them, and are able to state facts and a point of view, providing explanation and elaboration for their message. Their range of language and linguistic resources are such that they can engage and interact with the local speech community in common and unexceptional encounters, in person and online, using their L2. They have the propensity to do so even when they are being stretched linguistically.

Table 1. Level B1 descriptors as set out in the CEFR (Council of Europe, 2001).

Global scale (2001:24)	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.				
Self-assessment Grid (2001:26)	**Spoken interaction** I can deal with most situations likely to arise whilst travelling in an area where the language is spoken. I can enter unprepared into a conversation on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events).		**Spoken production** I can connect phrases in a simple way in order to describe experiences and events, my dreams, hopes and ambitions. I can briefly give reasons and explanations for opinions and plans. I can narrate a story or relate the plot of a book or film and describe my reactions.		
Qualitative aspects of Spoken Language Use (2001:29)	**Range** Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocution on topics such as family, hobbies and interests, work, travel and current events	**Accuracy** Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations.,	**Fluency** Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	**Interaction** Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual under-standing.	**Coherence** Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.
Phonological control (2001:117)	Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.				

3. The Relevance and Appropriateness of the CEFR B1 level

All language assessment frameworks are, to a greater or lesser extent, designed to respond to particular educational and/or social needs in specific contexts. So the CEFR B1 descriptors, indeed the full set of the CEFR descriptors, have to be understood with reference to the rationale underpinning their design in a Europe of the 1980s. One of the aims of the Council of Europe was (still is) to 'achieve greater unity among its members' and one way of achieving this was 'adopting common action in the cultural field' (Council of Europe, 2001:2). Teaching and learning the languages of the member states was part of the common action. Member states were encouraged to provide their citizens with 'a knowledge of the languages of other member states (or other communities within their own country, e.g. Italian speakers in the German-speaking region of South Tyrol in Italy)', and to use these languages to, inter alia, 'deal with the business of everyday life in another country, and to help foreigners staying in their own country to do so', 'to exchange information and ideas with young people and adults who speak a different language and to communicate their thoughts and feelings to them', and 'to achieve a wider and deeper understanding of the way of life and forms of thought of other peoples and of their cultural heritage' (op.cit., 2001:3). It was thought that xenophobia and ultra-nationalist sentiments were a 'primary obstacle to European mobility and integration', and 'a major threat to European stability and to the healthy functioning of democracy' (2001:4). (Although, some 35 years later, it is ironic that the CEFR is now used as a part of the screening apparatus for restricting migration and population mobility!)

Seen in this light, the CEFR projects an image of a European person from, say, Germany, speaking Italian in Italy as second or additional language as a foreign sojourner. The project of European unity and integration, in this view, promotes a pan-European platform for mobility of people (travel) and cultures (access to and understanding of different European national/regional cultures), with varied language proficiency in different European languages as a means of achieving this. There is strong imputation that such an individual is using their L2 as a sojourner visiting or staying in another (European) country, often only temporarily, for work or leisure.

4. Description Test Preparation Material

As indicated earlier, there are two approved tests, officially known as Secure English Language Tests: IELTS Life Skills (B1) and Trinity College London GESE Grade 5. Both are tests of listening and speaking (reading and writing are not assessed, although the Life in the UK test which is also mandated requires reading, yet it falls outside the scope of the present discussion). Although the two tests purportedly test similar skills at CEFR B1 level, they are somewhat different in nature.

4.1 IELTS Life Skills

The IELTS Life Skills is a test at which two candidates are present with an examiner. The test, which is in two parts, takes 22 minutes. It is made up of a number of components: a dialogue between the two candidates on a given topic including a question and answer segment; a monologue on a given topic, followed by a question and answer segment; a listening task during which candidates listen to a CD recording of two separate but related texts after which the examiner asks questions based on the messages heard; and a planning task that requires the two candidates to make joint choices (role play, for details see https://www.ielts.org/en-us/what-is-ielts/ielts-for-migration/united-kingdom/ielts-life-skills). Hereafter the items are described.

The first part of the test begins with the examiner asking each of the candidates in turn four prescribed questions eliciting the candidates' name, spelling of their name, where they are from and how long they have been living at their present location [see Example 1].

Example 1. (IELTS Part 1 Phase 1A).

Interviewer (I)	hello my names Sitar(.) what is your name
Candidate A (A)	my name's Imram
I	can you spell it for me
A	yes of course(.) its I m r a m
I	ok thank you(.) where do you come from Imram
A	I'm from Bangladesh
I	and how long have you lived here
A	I have lived here for 2 years

In the second phase of Part 12 candidates ask each other questions relating to a topic nominated by the examiner as shown in Example 2. In Phase 1B each candidate talks for one and a half minutes on another topic nominated by the examiner and once they have finished their mini-presentation the other candidate is told by the examiner to ask the speaker three questions related to the talk as in Example 3. Candidates have one minute to prepare for this phase of the test.

Example 2. (IELTS Part 1 Phase 1B).

Candidate A	(Turning to B) what is your favourite website
Candidate B	my favourite website is YouTube
A	why do you like this website
B	basically (.) YouTube is a video-sharing website (.) mostly the users upload different types of videos on this website (.) so I like watching educational videos shared on this website
A	how long have you been using it
B	I have been using it for around 2 years
A	how often do you use it

Example 3. (IELTS Part 1 Phase 1B).
[Imran has finished his mini-presentation and the interaction continues as follows]

I	thank you (.) Ayesha (.) please ask him then your 3 questions now
B	how long have you been friends for
A	we have been (uh) friends for 5 years
B	how do you stay in touch with each other
A	we (.) even we don't live nearby (uh) but we try (.) try to keep in touch through social media calls and texts
B	how often do you meet him
A	we often meet at the weekend

In Part 2 Phase 2A of the examination candidates listen twice to two short recordings after which the examiner asks them questions in turn. Before the recordings are played, the candidates are told by the examiner what specific information to listen for (see Example 4).

Example 4. (IELTS Part 2 Phase 2A).

I	thank you (.) now (.) listen to the information and answer these questions(.) Aysha in the first recording why can't Daniella go on the trip (.) why does Sally want to go to the park (.) Imran in the second recording (.) why are the tickets good (.) and where does he want to meet

In the final phase of the test (2B) the candidates perform two tasks. The first task involves planning something together. The candidates are told what they need to plan (e.g. a day out with a mutual friend) and the parameters of the plan (e.g. the options are: go shopping, visit historical places or go to the countryside). This task lasts two minutes after which the candidates may be interrupted before they have completed their plans. Example 5 is an extract from this task.

Example 5. (IELTS Part 2 Phase 2B).

A	well our friend is coming to visit our country this Sunday and he would like to spend that day with us (.) I'm really looking forward to meeting (.) to meeting him soon
B	I'm also eagerly waiting to meeting him (.) so what do you think we should plan to do together
A	I'm not sure what do you think
B	I was thinking of maybe visiting historical place
A	ah, that sounds like a good idea where do you think we should visit
B	well we can visit the Tower of London
A	that would be a good choice as he's coming to visit our country for the 1st time (.) he would definitely enjoy visiting that historical castle
B	also (.) he is like shopping so we could go shopping together (.) what do you think
A	no I think we should plan to spend at least 4 hours at the Tower of London as there is so much to see and learn about the place and its not possible to do both things on the same day
B	yes you are right (.) so lets stick to the idea of visiting the Tower of London only

The last task is similar to the one in Part 1 Phase 1A, in that the candidates exchange information about a further topic nominated by the examiner (e.g. how they spend their leisure time) but this time there is an element of comparison introduced into the task (e.g. tell your partner how you spend your leisure time now and how you spent it when you were younger) [Example 6].

Example 6. (IELTS Part 2 Phase 2C).

I	ok thank you (.) now you're going to talk together about free time activities so find out from each other what you used to do in the free time when you were younger and what you do now (.) so would you like to start Ayesha
B	yes sir [turning to B] tell me what you enjoy doing in your free time
A	well I usually watch tv in free time (.) mostly I like watching tele…. watching talk shows and (uh) sports programme
B	did you enjoy this activity when you were younger
A	no not at all (.) I used to enjoy outdoor activities like playing cricket or visiting new places when when I was younger though I still enjoy hanging out with my friends in my free time(.) so what do you enjoy doing in the free time
B	I love to read books whenever I get free time I read books
A	what sort of things do you like to read

The criteria upon which the candidates' performance is judged is based on the following: obtaining information, conveying information, speaking to communicate and engaging in discussion. Candidates are awarded a pass or fail on their performance (see https://www.ielts.org/-/media/publications/life-skills-guide-for-test-takers-and-agents/ielts-life-skills-guide-english-uk.ashx?la=en). Interestingly, no information is available in the public domain regarding what constitutes a pass.

4.2 Trinity College GESE Graded 5

The Trinity College GESE Grade 5 test, like the IELTS Life Skills test, is made up of two parts. However, unlike the IELTS equivalent, it only lasts about 10 minutes and the interaction is between an examiner and a single candidate. In part 1, the Topic Phase, the candidate chooses and prepares for a topic in advance and discusses the chosen topic with the examiner; in part 2, the Conversation Phase, the candidate engages in conversation on two topics chosen by the interviewer-examiner from a list of six options: festivals, transport, special occasions, entertainment, music, recent personal experiences (https://www.trinitycollege.com/site/?id=3365).

After brief introductions, in the first part of the test the examiner asks for the topic form on which the candidate has nominated a topic as well as 5 aspects of the said topic that the candidate wishes to discuss (e.g. My career: why I decided on this career; what I have achieved in my career so far; why my current job is preferable to my last job; what I was doing before I started to work in this job, and what I plan to do in the future). The interviewer-examiner then asks questions related to at least three of five aspects of the nominated topic. Example 7 is taken from the beginning of such an interaction.

Example 7. (GESE Grade 5).

I	ok you talk here about the differences here and your city where is your city?
C	in (.) in east Libya near the sea my city similar Brighton but there are some difference um (.) the weather (.) than here and um (.) small city no big and the building um very different uh.
I	I see how many people live in your city
C	40,000 about 40,000
I	I see a lot bigger I see ok now you say here you had some problems in Brighton
C	yes when I came here in Brighton (.) because my wife study I study and my wife study in another school. I have daughter 2 years uh (.) my problem in I don't I didn't find any nanny (.) for my wife.
I	for your baby

The second part of GESE Grade 5 is the conversation phase in which the examiner asks the candidate questions on two of the six prespecified subjects (see Example 8).

Example 8. (GESE Grade 5).

I C	I want to talk about entertainment (.) now eh what type of enter (…) do you prefer o stay at home for your entertainment or do you prefer to go out in the evening eh (.) I prefer to go with my friends out (.) (…) eh sometimes I stay at the home (…) watching the film (.) I prefer eh watch a film at home than the cinema home (…) watching the film (.) I prefer eh watch a film at home than the cinema
I	[latching] ah ha why why do you prefer it at home
C	mhm but at home relaxing and eh no noise (…) eh and thing
I C	eh ha OK have a cup of tea yeah with have a cup of coffee (a cup) and enjoy
I	and so will you go out this weekend
C	yes yes
I C	mhm mhm eh to the weather (…) very good (…) go out

Candidates are also required to ask the interviewer-examiner questions and although they can do so at any point of the exam, it would appear that more often than not, examiners ask candidates whether they have any questions for them as illustrated in Examples 9 & 10.

Example 9. (GESE Grade 5).

I	right thank you for telling me about that Aslam have you got a question for me about Brighton
C	what about you what about you think about living here in Brighton

Example 10. (GESE Grade 5).

I	have you got a question for me either about special occasions or about entertainment
C	would you prefer watching film in the home or go to the cinema or …. [latching] eh it depends

Candidates can pass at one of three levels: A, B or C (with D constituting a fail) and information is publicly available explaining these levels as well as showing examples of the test in action.

For reasons of focus and scope, our specific analytic attention is on the speech functions involved in the interactional talk between the examiner and the candidate and between the candidates themselves. While the listening comprehension tasks are an important component of the IELTS test, taken as a whole interactional talk constitutes a substantial part of the two speaking tests. Furthermore, interactional talk is likely to be a significant part of the candidates' ability to handle their communication needs as an active participant in community and work contexts.

5. Analytic Framework

Our primary analytic interest in examining the interactional discourse in the training materials is to identify the speech functions, as represented by the interactional moves, in the talk as the test unfolds. We understand that the video material that is publicly available is there for training purposes, so it may not be authentic test data. However, we would suggest that as officially approved training material the videos are a significant presentation of what the tests comprise in terms of expected interactions and the associated use of language. Conceptually we would argue that the training videos are manifestations of the underlying constructs.

We draw on the discourse analytic frame proposed by Eggins & Slade (1997) which is designed to make visible the interactional options (expressed as moves) made by the interlocutors, in this case the interviewer-examiner and the candidate/s. Figure 1 provides a schematic representation of some of the interactional moves that will help us frame the description of the talk between the examiners and the candidate/s with a view to showing conversational work being enacted.

In the Eggins and Slade framework, spoken interaction is primarily seen as comprising two initial moves: Open and Sustain. An Open move sets up an exchange either at the beginning of a conversation, or in the midst of a conversation to trigger a new exchange on a different topic. It can serve two functions: (a) setting the scene for interaction (Attend) by drawing attention to and acknowledging the immediate context of the interaction through salutations, greeting and invocations of names, which can 'prepare the ground for interaction by securing the attention of the intended interactant'; or (b) 'getting the interaction underway' by keying the conversation partner/s into a request or demand for information or goods and services', i.e. to [I]nitiate a topic (Eggins and Slade, 1997:193).

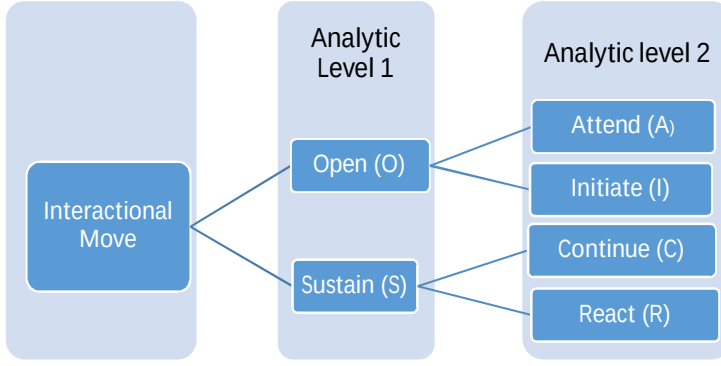


Figure 1. Interactional Moves (adapted from Eggins and Slade, 1997:193, 195, 202, 209).

A Sustain move responds to an Open move. Sustain moves, on the part of a conversation partner, can serve two functions: (a) maintaining and extending the topic nominated in the Open move by, for example, checking and trying to verify the speaker's own understanding of the information in the on-going talk or by providing further information to [C]ontinue with the Initiate, or (b) [R]eacting to the Open move by negotiating, extending and completing an idea or a proposal on the terms set up by a previous speaker or by rejecting or querying the proposition or content in the Open move.

It should be pointed out that the Eggins and Slade framework has a 5-level hierarchy containing more analytic delicacy than what is shown here. For instance, the Continue move is further expanded as Monitor, Prolong and Append; and both Prolong and Append are further instantiated as Elaborate, Extend and Enhance. For the purpose of this discussion though the Open and Sustain moves (analytic level 1) and the Attend, Initiate, Continue and React moves (analytic level 2) are the most helpful in framing the interactional moves.

The interactional moves are expressed in utterances encoded in grammatical forms such as declarative statements and interrogatives, but there is no necessary one-to-one correspondence between grammatical categories and speech functions in actual discourse. In other words, in conversation discourse meaning cannot be read off by looking at grammatical forms. For instance, an interrogative such as 'Did you say you don't like watching television?' is not necessarily intended to be understood as a question soliciting confirmation of what has been said. In an actual interactional context it is possible that this utterance is intended to be understood as an invitation to elaborate on a point already made. So grammatical form and discourse functions are not always congruent. The Eggins and Slade analytic framework offers a discourse semantic network of options representing speech functions that can be used to help track the moves made by the examiners and candidates.

We do not suggest that the options of speech functions presented in the schematic diagram above is exhaustive. Interlocutors, including candidates in test situations, may engage in conversation in agentive and novel ways. From an analytic point of view, the semantic options framework we adopted will help us show what kind of conversation 'work' is being done in the training videos.

6. The Sample Test Discourse

An analysis of the discourse elicited during the sample tests described above highlights a number of features specific to test performance.

(a) The nature of the interviewer questions

In both tests the interviewer-examiner asks a number of predictable questions that elicit from candidates what can best be described as question-answer ping-pong with very formulaic turn-taking. Example 1 from the beginning of the IELTS interview shows, using the Eggins and Slade (1997) framework explained above, in order to sustain the interaction, each of the participants follows a set pattern in which the speech function use is not reciprocal among participants. (In all the following examples, the moves are placed to the left in order to highlight these, while the interactant is placed on the right.)

Example 11.

O- open; R – react

Move	**Language expression**	**Speaker**
O-attend -initiates interaction by asking question R-responds to question	hello my name's sitar what's your name my name's imram	I A
R-asks for more information	can you spell it for me	I
R-responds to invitation R-responds to question	yes of course it's i-m-r-a-n	A
R-acknowledges reply R-asks for further information	ok where do you come from imram	I

A similarly restrictive pattern of interaction is evident through much of the Trinity material. Although in the example below there is some extension/elaboration in the first of the candidate's responses, the lack of engagement in this interaction is evident from the examiner's final turn where the examiner abruptly switches topic rather than extend the discussion showing interest in the candidate's home town.

Example 12.

Move	**Language expression**	**Speaker**
O-attend -initiates interaction by asking for information	ok you talk here about the differences here and your city where is your city	I
R-responds to question C*-adds additional information C-elaborates further on reply	in (.) in east Libya near the sea similar Brighton but there are some difference um (.) the weather (.) than here and um (.) small city no big and the building um very different uh	Candidate
R-acknowledges reply C-asks for further information	I see how many people live in your city	I
R-responds to question	40,000 about 40,000	Candidate
R-acknowledges response R-leads into new question	I see a lot bigger I see ok now you say here you had some problems in Brighton	I

C* where C = Continue

Where the candidate asks a question in the Trinity exam, this again does not arise from genuine interest, but from an obligation to display an ability to form grammatically correct questions, as in Examples 9 and 10. What is also worth noting here is that the candidate tends to echo the examiner's questions.

During the IELTS exam where candidates ask each other questions, however, the questions do not seem to arise from genuine interest but a need to comply with the interviewer-examiner's instructions. This is evident in the first turn of Example 3 (repeated below).

Example 13. (repeated from Example 3).

Move	**Language expression**	**Speaker**
R- acknowledges (rubric driven examiner talk)	thank you ayesha, please ask him then your 3 questions now	I
R – responds to examiner's request, but with question for A	how long have you been friends for?	B
R-responds to question	we have been (uh) friends for 5 years	A
R-asks 2nd question	how do you stay in touch with each other?	B
R-responds to 2nd question	we, even we don't live nearby (uh) but we try … try to keep in touch through social media calls and texts	A
R-asks 3rd question	how often do you meet him?	B
R-responds to 3rd question	we often meet at the weekend.	A

The first two turns in this part of the interaction are difficult to analyse in terms of casual conversation as the opening move is fulfilled by the examiner (I) and Candidate A then asks the first question to Candidate B not the examiner, so this turn can neither be viewed as an opening move nor a sustaining one. The proceeding interaction then follows a set pattern (similar to that in Example 1) where one person asks the questions and the other responds. Again the limited extension and elaboration of the discourse is evident here, yet at the same time Candidate A's responses appear very complete (and absent of ellipses). In response to the question 'How often do you meet him?', Candidate A could have answered 'most weekends' rather than using 'We often meet at weekends'. The latter more complete form appears pervasive throughout the analysed tests and appears more often associated with test language rather than normal conversational language as described by Eggins and Slade (1997). It is possible, indeed likely, that this is an interpretation of the CEFR Level B1 qualities in respect of Range, Interaction and Coherence (see Table 1).

(b) Reacting moves

An important point to note is that throughout both tests reacting moves are predominantly supportive. In the data analysed, there is only one instance of a confronting move which occurs during the second phase of the IELTS Life Skills test (illustrated below). And, even here, as soon as Candidate A does not agree with Candidate B, the latter immediately complies (turn 4).

Example 14. (extract from Example 5).

Move	**Language expression**	**Speaker**
C-extends response by adding further information	**(.) he would definitely enjoy visiting that historical castle.**	**A**
R-puts forward alternative idea	also he is like shopping so we could go shopping together	B
R-probes opinion	what do you think?	
R-responds by disagreeing	no	A
C-extends response	I think we should plan to spend at least 4 hours at the Tower of London	
R-responds	yes you are right (.)	B
C-extends response	so lets stick to the idea of visiting the Tower of London only	

Where the interaction is between a candidate and examiner, the power-relationship between the speakers would tend to favour compliance rather than confrontation. But even when the talk is among candidates themselves, there would appear to be little opportunity to confront what the speaker is saying and doing so is often specifically discouraged when preparing candidates for such tests (see, e.g., https://www.youtube. com/ watch? v=s3 ErfZVIM3c).

(c) Opening moves

As has been implied above, the opening moves of test discourse are difficult to analyse as these are inevitably initiated by the examiner. The role of the examiner as the initiator of any interaction invariably impacts on the social relationships that are established during the test and the speech behaviour that subsequently ensues. In the Trinity GESE exam, even when the topic of the discourse has been nominated by the candidate (Part 1 of the test), the interaction is opened and driven by the examiner. And, as has been shown above, where the candidates are to talk among themselves in the IELTS Life Skills test, they do not initiate the talk: they wait for the examiner to do so. Just like the candidates in the GESE exam, their interaction takes place within a very limited range of topics that to a large extent can be predicted. The sample test materials also strongly suggest that the interactional talk is highly constrained.

(d) 'Bulge' language

In situations where social relationships are public but their maintenance is open to negotiation, people tend to want to signal as much social solidarity as possible and to avoid conflict or confrontation. This is precisely the kind of 'polite' and non-offensive language that is regarded as suitable for use in the public domain and under the gaze of others (e.g. an examiner's scrutiny). Wolfson (1986) refers to this kind of language as 'the bulge' (also see Cook, 2000 and Coffey and Leung, 2019, aop). In contexts where social relationship is more clearly known or defined (by others) – i.e. above and below

the ‘bulge’ – the range of variegated language in terms of directness, endearment and rudeness is likely to be far greater.

‘Bulge’ language tends to occur where participant roles are reasonably clear to all involved, but the actual language use is subject to social or other kinds of evaluation. In this case, the candidates’ language use is subject to the raters’ assessments. The tendency to be driven by and to converge on the ‘polite’ and, relatively speaking, personally non-involved language in test-influenced examiner-candidate exchanges has implications for validity claims. A key question here is whether ‘bulge’ language should form a proficiency benchmark suitable for citizenship. We will explore this question next.

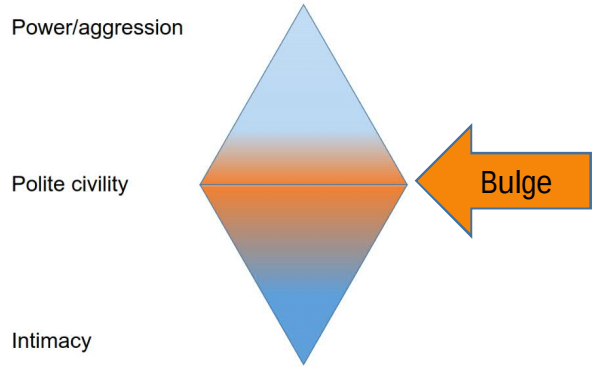


Figure 2. Bulge Language (Coffey and Leung, 2009, aop).

7. Language of the New Citizen

The idea that new citizens, through migration and residency, should speak the national or the official language(s) of their chosen home country has now been incorporated into legislation in many countries. The requirement for new citizens to demonstrate their language proficiency as part of the qualifications for granting residency is well established in many countries in Europe, Australia and the USA. In the case of the UK, the association between English language proficiency and citizenship (a part of which is the granting of permanent residency) was first mooted over 15 years ago. David Blunkett, the then British Home Secretary, was one of the first to draw official attention to the issue of English language proficiency for economic and social participation for migrants in 2001 (https:// www.telegraph.co.uk/news/uknews/ 1337901/Migrants-must -learn-English-in-Blunkett-plan.html). One strand of the political opinion was that the lack of English proficiency on the part of some members of ethnic minority

communities was a major contributing factor to the perceived lack of social cohesion in British society; it was also argued that the lack of English was holding back people belonging to minority communities from economic participation and, therefore, was reducing their opportunities to advance their life chances in society. Policy and legal changes in line with the view that the ability to speak English was a requisite for greater national cohesion followed from 2002 onwards (see Blackledge, 2009). However, the issue has not disappeared from public debate despite the fact that English language proficiency is now one of the statutory requirements included in the application for permanent residency and naturalization. The violent acts perpetrated by some minority religious groups across Europe and elsewhere in recent times have re-energised the call for social integration and cohesion. Very recently, Dame Louise Casey, a high profile national political figure and the author of an influential parliamentary report on social cohesion and community integration, called for legislation to make sure that English is learned by everyone with a deadline: 'I would set a target that says by X date we want everybody in the country to be able to speak a common language' (https:// www. independent.co.uk/news/uk/home-news/uk-speak-english-language-integration-tsar-dame-louise-casey-immigrants-a8252311.html). This continuing political anxiety about the lack of English among the migrant and ethnic minority communities has been officially picked up in a 2018 government policy paper 'Integrated Communities Strategy' in which Sajid Javid, the current Home Secretary, announces that '£50 million will be committed to over the next 2 years' to, inter alia, 'boost English language skills' and to 'promote British values' (https://www.gov.uk/government/news/new-government-action-to-create-stronger-more-integrated-britain).

The relationship between language and citizenship is complex and highly contingent on local context; there are many different ways of conceptualizing language(s) and uses of language(s) in society (e.g. Williams & Stroud, 2015; Rampton, Cooke & Holmes, 2018). The official British government discourse on language and social cohesion is of course a contested ideological articulation (see Blackledge, 2009; Bruzos, Erdocia & Khan, 2017; Cooke, 2009; McNamara, Khan & Frost, 2015; Simpson & Whiteside, 2015). It seems to be premised on the idea that the English language is an embodiment of a set of preferred British values (however defined), and that all speakers of the language would subscribe to and practise these values. There is little prima facie evidence to suggest that both sharing and speaking the same language automatically lead to universal subscription to a set of common values and greater social cohesion. From an international perspective, for instance, the sharing of Castillian Spanish as a national language in Spain does not seem to have reduced the demands for regional separatism, as recent events would readily attest. Nationally it does not require any specialist knowledge to understand that a penal code is needed in the UK (as in virtually all societies) to deal with anti-social behaviour, despite the fact that English is the shared language for the vast majority of the people. At a more local (area by area) level, the often assumed direct link between ethnic/racial diversity and a lack of social cohesion, or the lack of ability to speak a shared language and committing

crime(s) is not borne out by empirical evidence (e.g. Letki, 2008; Gostjev & Nielsen, 2007). So it is clear that the putative causal link between language and social cohesion is clearly in need of further exploration and validation. The usefulness of being able to speak the national language(s) for personal, educational, professional and social purposes is, however, much less controversial. Given the focus and scope of this chapter we will not dwell further on the constitutional, legal and moral relationship between language and citizenship. Instead, we will now look at the question of what sorts of language knowledge and skills new citizens may need to go about their everyday lives in society.

The juxtaposition of the need for social integration and English language signals an assumption that some members of migrant and minority communities lead lives that are outside the mainstream society (however defined), and their lack of English is making fuller participation in mainstream society difficult, if not impossible. The projected imagery is that of (groups of) individuals from ethnic minority backgrounds living in small separate communities completely insulated from the English-speaking national culture and practices. While this imagery has gained traction in political rhetoric, the question here is: How valid is this portrayal? While there is no doubt that identifiably 'ethnic' communities do exist - by virtue of their demographics and religious-cultural practices – it is by no means clear how they can exist in some form of splendid isolation from the rest of society. Some time ago, Hall (2000:221) observed that in contemporary diverse Britain people live through multiple cultures and communities. A realistic understanding of the diverse and complex intersections between the different communities would need to take account of the

> . […] lived complexity emerging in these diasporic communities, where the so-called "traditional" ways of life derived from the cultures of origin remain important to community self-definitions, but consistently operate alongside daily interaction at every level, with British mainstream life.

This view is echoed by Gilroy (2004) who suggests that there is a kind of convivial rubbing along among members of diverse communities in everyday life (against the backdrop of some vestiges of empire and colonialism). All of this suggests that there is a need to look more closely at what goes on at the ground level.

At the level of policy rhetoric, the projected Andersonian 'imagined' idealised national community privileges a top-down view that foregrounds the importance of a common national culture and identity (see Anderson, 1991). However, this projected common national culture and identity is premised on a somewhat top-down over-simplified narrative of the lived experiences of citizens; it offers a partial depiction at best. At the level of everyday life Alexander, Edwards and Temple (2007:788) argue that people live in, and traverse through, their personal communities: 'Rather than being an abstract category, 'community' is lived through embedded networks of individual, family and group histories, trajectories and experiences'. For the individual, then, their personal community comprises 'local, heterogeneous and contingent

networks of family, friends and neighbours linked and performed through ties of emotion, trust and security' (Alexander et al., 2007). As Isin (2009) observes, denizens (whatever their legal residential status) in contemporary societies unavoidably come into contact with and have to navigate through a large number of societal sites associated with the conduct of the globalized economy and industrial production, local and international legal entities, as well as civic and political organisations. In traversing such complex networks of connections in their everyday activities everyone in society (including aspirant future citizens) inevitably has to engage with a multitude of civic, commercial, cultural, professional and religious organisations, as well as local and national governmental agencies, all with their social and institutional norms and practices. The language repertoire needed for effective communication in these encounters is unlikely to be limited to that sampled by the tests discussed earlier. It is through this perspective that we can begin to understand the relationship between language and citizenship more realistically.

Through the lens of personal community, it is now possible to develop a (more) close-up picture of new citizens' (indeed any citizen's) everyday activities. At a minimum we can say a personal community would, inter alia, comprise family and local community networks; welfare agencies (e.g. housing and social welfare) and public services (e.g. health services), educational and professional institutions; as well as religious, sporting and other lifestyle organisations. Individuals participate in, interact with and traverse through these groupings, organisations and institutions in various capacities at different points of time for different purposes. The language used in this complex web of activities would necessarily involve a wide range of registers, styles and functions, from the routinely transactional to the highly intimate and emotional, from the business-like to the aesthetically tuned. And, of course, some parts of the personal community would likely take place using language(s) other than English; multilingualism is likely to be a significant feature of personal communities. The salient question now is: Does the content of the officially sanctioned secure spoken language test at CEFR Level B1 speak to the language communicative needs and uses of the new citizen?

8. Concluding Remarks

The discourse elicited during both of the speaking sample tests examined appears very restricted both in terms of topic and content as well as the discourse moves. Candidates do not initiate (open) discussion, they do not engage in agentive talk (always enacting within the parameters of test talk) and rarely, if ever, have the chance to confront or challenge what their interlocutors say. The language fits the pattern of 'unemotional everyday public transactions' (Cook, 2000:62). It does not provide opportunities for agentive use of language which raises the all-important question of whether the tests are fit for purpose. The construct being tested appears to be limited and thus under-represents the idea of social participation in the community identified by the Home Office who has mandated the test.

While we have not seen any official statement of the construct of the two secure tests, it seems reasonable to suggest that it is underpinned by the CEFR descriptors at B1 level. As mentioned earlier, CEFR B1 speaking is highly suggestive of the sojourner within a European context. CEFR Level B1 speaking would seem to us to be about communicating with others in the capacity of a visitor or traveller and yet the tests are now being used to assess the speaking abilities of new residents and citizens who are expected to be fully socially integrated so as to participate fully in personal, educational, professional and other community activities in an agentive manner.

This clearly raises the issue of fitness for purpose and, therefore, we would very much like to see further discussion on the appropriate construct for this assessment. It also raises the wider question of whether this form of language assessment can ever adequately cover the full range of language use(s) as a citizen. After all, we do not ask our UK born citizens to demonstrate their speaking abilities as they transverse through life in society at different ages. Within the UK there are, for example, speakers of Welsh and Gaelic who may have limited fluency in English. Should there be an English language assessment issue arising from this? There is thus an issue of fairness of equal treatment for all residents that cannot be ignored.

We fully understand that the rating scales within the CEFR are intended to be used as 'a tool to facilitate education reform projects, not a standardisation tool' (Council of Europe, 2018:26). Nevertheless, it is also claimed that the 'CEFR scales are intended to be used to profile ability' (2018:53). Given all the conceptual and practical problems as well as challenges we (and many others) have identified, it would seem that at this stage we are left with a number of fundamental questions regarding adequate profiling, including:

1. Assuming that some form of assessment of language proficiency for aspirant new citizens is here to stay, is CEFR Level B1 speaking appropriate for capturing the language needs of permanents residents?
2. Related to the question above, what kind(s) of test or assessment design and format would be appropriate and suitable for the purpose at hand?
3. Is it possible to capture the language needs of any resident at different points in their lives?
4. Even if it were possible to capture these needs, would it be fair to assess only one group of residents?

We leave you, the reader, to reflect on the above as these are just some of the questions which need to be addressed if we are to begin to understand what counts as language proficiency in relation to citizenship.

References

Alexander, C., Edwards, R. & Temple, B. (2007). Contesting cultural communities: Language, ethnicity and citizenship in Britain. *Journal of Ethnic and Migration Studies*, 33(5), 783-800.

Anderson, B. (1991). *Imagined Communities: Reflections on the Origin and Spread of Nationalism* (Rev. ed.). London: Verso.

Blackledge, A. (2009). "As a country we do expect": The further extension of language testing regimes in the United Kingdom. *Language Assessment Quarterly*, 6(1), 6-16.

Bruzos, A., Erdocia, I. & Khan, K. (2017). The path to naturalization in Spain: Old ideologies, new language testing regimes and the problem of test use. *Language Policy*, First Online. Retrieved from Coffey, S., & Leung, C. (2019). Understanding agency and constraints in the conception of creativity in the language classroom. *Applied Linguistic Review, Advance online publication*.

Coffey, S. & Leung, C. (2019). Understanding agency and constraints in the conception of creativity in the language classroom. *Applied Linguistic Review, Advance online publication*.

Cook, G. (2000). *Language play, language learning.* Oxford: Oxford University Press.

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

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning teaching, assessment* - Companion Volume with New Descriptors. Strasbourg: Council of Europe.

Eggins, S. & Slade, D. (1997). *Analysing casual conversation*. London: Cassell.

Extra, G., Spotti, M., & Van Avermaet, P. (2009). Testing regimes for newcomers. In G. Extra, M. Spotti & P. Van Avermaet (Eds.), *Language testing, migration and citizenship: Cross-national perspectives on integration regimes* (pp. 1-33). London: Continuum.

Gilroy, P. (2004). A*fter empire: Melancholia or convivial culture? Multiculture or postcolonial melancholia.* London: Routledge.

Gostjev, F. A. & Nielsen, A. L. (2016). Speaking the same language? English Language fluency and violent crime at the neighborhood level. *The Sociological Quarterly*, 58(1), 111-139.

Gysen, S., Kuijper, H. & Van Avermaet, P. (2009). Language testing in the context of immigration and citizenship: The case of the Netherlands and Flanders (Belgium). *Language Assessment Quarterly*, 6(1), 98-105.

Hall, S. (2000). The multicultural question. In B. Hesse (Ed.), *Un/settled multiculuralisms* (pp. 209-241). London: Zed Press.

Isin, E. F. (2009). Citizenship in flux: The figure of the activist citizen. *Subjectivity*, 29(1), 367-388.

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, 27(2), 183-189.

Letki, N. (2008). Does diversity erode social cohesion? Social capital and race in British neighbourhoods. *Political Studies*, 56(1), 99-126.

McNamara, T. (2005). 21st century shibboleth: Language tests, identity and intergroup conflict. *Language Policy*, 4(4), 351-370.

McNamara, T., Khan, K. & Frost, K. (2015). Language tests for residency and citizenship and the conferring of individuality. In B. Spolsky, O. Inbar-Lourie & M. Tannenbaum (Eds.), *Challenges for language education and policy* (pp. 11-22). London: Routledge.

Rampton, B., Cooke, M. & Holmes, S. (2018). Promoting linguistic citizenship: Issues, problems and possibilities. In *Working Papers in Urban Language and Literacies* (pp. 1-29): University of Gent, State University of New York at Albany, Tilburg University, King's College London.

Simpson, J. & Whiteside, A. (Eds.). (2015). *Adult language education and migration: Challenging agendas in policy and practice*. London: Routledge.

Webster, W. (2018). Windrush generation: The history of unbelonging. *The Conversation* 2018-04-18.

Williams, Q. E. & Stroud, C. (2015). Linguistic citizenship: Language and politics in postnational modernities. *Journal of Language and Politics*, 14(3), 406-430.

Wolfson, N. (1986). The Bulge: A theory of speech behavior and social distance. *Working Papers in Educational Linguistics* (UPenn), 55-83.

Plurilingual and intercultural education: Some critical reflections

David Little
Trinity College Dublin

1. Introduction

Since the publication of the *Common European Framework of Reference for Languages* (CEFR, 2001), the Council of Europe's work on language education policy and practice has been shaped by the concepts of plurilingualism and interculturality, yoked together in the doctrine of "plurilingual and intercultural education". In this article I offer some critical reflections on this new orthodoxy, which seems to me seriously vulnerable on theoretical and empirical grounds. I begin by analysing the CEFR's foundational definition of "the plurilingual approach", pointing out its contradictions and considering its implications for second language pedagogy. I then turn to "pluriculturalism", which the CEFR offers as plurilingualism's twin, and "interculturality", the concept that replaced it in the Council of Europe's project *Languages in Education, Languages for Education*. In the third part of the article I describe two language learning environments with which I am closely familiar to show that the development of fully integrated plurilingual repertoires in contexts of formal education is entirely possible, but that it entails the adoption of pedagogical approaches that focus only incidentally on otherness and cultural difference. In other words, I question whether the link between the development of plurilingual repertoires and intercultural education is inevitable or necessary. The CEFR itself concedes that the full implications of its plurilingual approach have still to be worked out and translated into practice (CEFR, 2001:5). Coming seventeen years after the first publication of the CEFR's, this article is a long overdue contribution to that process.

2. The concept of plurilingualism and its implications for second language pedagogy

The foundational definition of the Council of Europe's "plurilingual approach" is given in the first chapter of the CEFR (2001: 4):

> Plurilingualism differs from multilingualism, which is the knowledge of a number of languages, or the co-existence of different languages in a given society. Multilingualism may be attained by simply diversifying the languages on offer in a particular school or educational system, or by encouraging pupils to learn more than one foreign language, or reducing the dominant position of English in international communication. Beyond this, the plurilingual approach emphasises the fact that as an individual person's experience of language in its cultural contexts expands, from the language of the home to that of society at large and then to the languages of other peoples (whether learnt at school or college, or by direct experience), he or she does not keep those languages and cultures in strictly separated mental compartments, but rather builds up a communicative competence to which all knowledge and experience of languages contributes, and in which languages interrelate and interact.

The essence of this definition comes at the end: plurilingualism is "a communicative competence to which all knowledge and experience of languages contributes, and in which languages interrelate and interact". The intended meaning of the verbs "interrelate" and "interact" is unclear, but the strong implication is that in appropriate contexts, all languages in a plurilingual repertoire are equally available for immediate, spontaneous use. The phrase "in appropriate contexts" accommodates two facts. First, language users/learners are typically stronger in some communicative modes than in others; they may, for example, be stronger in listening and speaking than in reading and writing. Second, it is quite usual for the different languages in a plurilingual repertoire to come into their own in different contexts; for example, one language may rarely be used outside the home, remaining relatively underdeveloped in more formal registers, whereas educational success presupposes the development of high levels of specialized literacy in the language of schooling. Understood in this way, plurilingualism is clearly a close relative of linguistic multicompetence, originally defined by Vivian Cook (1991:112) as "the compound state of a mind with two [or more] grammars". Both definitions imply that each language in a plurilingual repertoire is part of the user/learner's "everyday lived language" (García, 2017:18): rooted in his or her identity, a part of what he or she is, and a channel of his or her agency.

It is one thing to adopt plurilingualism as an appropriate goal for language education, quite another to decide how to achieve it. Unfortunately, instead of providing us with useful pointers, the CEFR's foundational definition entangles itself in contradiction. Since the publication of the CEFR, the Council of Europe has mostly applied plurilingualism to the individual citizen and multilingualism to communities and societies. This is a useful distinction, at least in those languages whose lexicon admits the pluri/multi contrast. After all, many plurilingual individuals live in predominantly monolingual societies, and some multilingual societies include large numbers of monolingual individuals. But in defining plurilingualism, the CEFR also applies multilingualism to the individual language user/learner. Besides referring to "the co-existence of different languages in a given society", we are told (CEFR,2001:4), multilingualism is also "the knowledge of a number of languages", which

> may be attained by simply diversifying the languages on offer in a particular school or educational system, or by encouraging pupils to learn more than one foreign language, or reducing the dominant position of English in international communication.

This is evidently a twofold criticism, of the tradition of teaching languages in isolation from one another and of the tendency in non-anglophone countries for English to be taught to the exclusion of other languages. The implied claim that there is a psycholinguistic difference between plurilingualism and individual multilingualism is clearly absurd: there is no evidence to suggest that the way in which languages are arranged in the curriculum and taught in classrooms determines how they are stored in the mind and accessed for use. Nevertheless, the fractured status of individual multilingualism compared with the integrated communicative competence of plurilingualism invites the conclusion that the plurilingual approach requires innovation in pedagogy as well as curricula.

The next part of the text, however, effectively undermines whatever may have been gained by contrasting individual multilingualism with plurilingualism. Plurilingual repertoires develop, we are told (CEFR, 2001:4):

> as an individual person's experience of language in its cultural contexts expands, from the language of the home to that of society at large and then to the languages of other peoples (whether learnt at school or college, or by direct experience).

Plurilingual repertoires, in other words, are simply the result of learning multiple languages, regardless of the context in which learning takes place and the dynamic by which it proceeds. At this point in the definition we are asked to believe that there is no qualitative difference between the processes and outcomes of child language development, naturalistic second language acquisition, and language learning at school. This would be an excusable short-cut if languages learnt at school routinely took their place in learners' plurilingual repertoires, being available for immediate and spontaneous use in appropriate contexts, but they are not, and never have been. Recognition of this fact was what led the Council of Europe to develop new ways of defining language learning goals in the 1970s, to promote communicative approaches to language teaching in the 1980s, and to develop the CEFR itself in the 1990s. Despite these initiatives and accompanying Recommendations from the Committee of Ministers to Council of Europe member states, language learning outcomes remain stubbornly disappointing in many countries. Empirical confirmation of this fact was provided by the European Commission's First European Survey on Language Competences (ESLC, 2011) carried out in sixteen EU countries and regions, and which focused on learners at, or approaching, the end of compulsory schooling.

The ESLC results show an overall low level of competences in both first and second foreign languages tested. The level of independent user (B1/B2) is achieved by only 42% of tested students in the first foreign language, and by only 25% in the second foreign language. Moreover, a large number of pupils did not even achieve the level of

a basic user: 14% for the first and 20 % for the second foreign language (European Commission, 2012:5).

On the basis of my analysis so far, the CEFR's foundational definition of its plurilingual approach may be summarized as follows. Plurilingualism is "a communicative competence to which all knowledge and experience of languages contributes"; it differs from individual multilingualism, which comes from learning multiple languages in isolation from one another – normal practice in most education systems. This difference may be thought to imply that the widespread achievement of plurilingualism in formal education will require innovation in curriculum and pedagogy. At the same time, however, the CEFR makes no distinction between developmental, experiential and instructed language learning. Confusingly, this implies that plurilingualism, like individual multilingualism, is a matter simply of learning multiple languages; so perhaps nothing needs to change after all. At this point it is appropriate to switch the focus to pluriculturalism and interculturality.

3. Pluricultural competence and interculturality

It is easy to overlook the single mention of culture in the CEFR's definition of the plurilingual approach (2001:4):

> as an individual person's experience of language in its *cultural* contexts expands, from the language of the home to that of society at large and then to the languages of other peoples (whether learnt at school or college, or by direct experience) (emphasis added).

I have already pointed out that these words elide the many differences between child language development, naturalistic second language acquisition, and instructed language learning. As small children acquire the language of their immediate environment they also acquire its culture, understood in the widest possible sense: behaviours, attitudes, beliefs, and intimate familiarity with a wide range of cultural artefacts. And because naturalistic second language acquisition is the result of living among speakers of the target language, it is likely to include the acquisition of culture, though to a more limited extent: what the learner acquires of the host culture will be a function of the extent to which he or she shares in the life of the community in question. But in what sense can we say that foreign languages are taught and learnt "in their cultural context" at school? Does the CEFR represent the essentialist view that each language comes with its own culture somehow included? If so, by what features are we to recognize that culture?

Later in the CEFR (2001:168) plurilingual and pluricultural competence are presented as a unity, two sides of the same coin:

> Plurilingual and pluricultural competence refers to the ability to use languages for the purposes of communication and to take part in intercultural interaction, where a person viewed as a social agent has proficiency, of varying degrees, in several languages and experience of several cultures.

Again, there is no difficulty in accepting this view when it is applied to languages that are acquired developmentally or naturalistically. But what about languages that are learnt as part of formal education? The CEFR concedes that there are not necessarily any "links between the development of abilities concerned with relating to other cultures and the development of linguistic communicative proficiency" (ibid.). This seems to be aimed, however, at those who insist that pluricultural competence is not exclusively linguistic; it does not throw light on the relation between language and culture.

A few years after the publication of the CEFR, the Council of Europe's project *Languages in Education, Languages for Education* adopted "intercultural" in place of "pluricultural". The rationale for this change is provided in a short text by Michael Byram. Arguing against a simplistic view of the relation between language and culture, Byram defines culture as "something established, belonging to a particular national, ethnic, religious or other 'community', and as a dynamic process relying on personal choice" (Byram, 2009:5). In other words, culture is at once objective and subjective: something about which we can learn, possibly at a distance, and something that we acquire from personal involvement and experience. This latter dimension is fundamental to Byram's definition of pluriculturalism (2009:6), which:

> involves identifying with at least some of the values, beliefs and/or practices of two or more cultures, as well as acquiring the competences which are necessary for actively participating in those cultures.

According to this definition, pluricultural individuals:

> are more likely to come from ethnic minority than ethnic majority backgrounds, because minority individuals usually have not [only] their own ethnic heritage culture but must also engage with aspects of the dominant majority national culture in which they live (ibid.).

Plurilingualism is the natural result of living one's life through two or more languages, while pluriculturalism arises from living in two or more cultures. If we adopt plurilingualism as the goal of language education, the challenge we face is to find ways of ensuring that, on however modest a scale, language learners at school and college live their lives through the languages they are learning. The simultaneous development of pluriculturalism, however, is a logical impossibility. It is true that the acquisition of increasingly advanced proficiency in the language of schooling entails acquisition of a series of academic sub-cultures, as does the use of a foreign language in CLIL projects; but the same cannot be said of languages that are included in the curriculum for their

own sake. We can teach second and foreign languages in ways that take account of culture in Byram's objective sense, but we cannot provide students with a version of the subjective experience of culture that accompanies developmental and naturalistic language acquisition.

These considerations explain why the project *Languages in Education/ Languages for Education* decided to replace pluriculturality, "the capacity to identify with and participate in multiple cultures", by interculturality (Byram, 2009:6):

> the capacity to experience and analyse cultural otherness, and to use this experience to reflect on matters that are usually taken for granted within one's own culture and environment.

The relation between plurilingual and intercultural is clearly more relaxed than the relation between plurilingual and pluricultural. The cultural element in language education loses its tinge of essentialism and becomes a matter of developing a "multiperspectivity" to which different languages and different areas of the curriculum can contribute in different ways (Byram, 2009:9):

> The panoply of languages which consists of regional, minority and migration languages, the language(s) of schooling and foreign languages is the means for expressing these perspectives, and the modes of teaching these languages need to take this into account in various ways, in particular comparing and contrasting perspectives on "the same" phenomena.

These words prompt two questions. First, what "modes of teaching" are available if we are to promote the development of plurilingual repertoires in the CEFR's core sense? And second, are those modes easily compatible with the exploration and expression of multiple cultural perspectives? In order to answer these questions, it is necessary to switch our attention from theory to practice.

4. Developing plurilingual repertoires at school: Two examples

In Chapter 6 of the CEFR (2001:142) we are told that "it is not the function of the Framework to promote one particular language teaching methodology, but instead to present options". These words have sometimes been used to argue that the CEFR is methodologically neutral. They are, however, preceded by the following reminder (ibid.):

> For many years the Council of Europe has promoted an approach [to language teaching] based on the communicative needs of learners and the use of materials and methods that will enable learners to satisfy these needs and which are appropriate to their characteristics as learners.

The fact is, moreover, that the CEFR's description of language proficiency in terms of language use carries the inescapable implication that target language use should play a central role in teaching and learning. But what kind of target language use? We can deduce the beginnings of an answer to this question from the CEFR's description of its action-oriented approach (2001:9):

> The approach adopted here, generally speaking, is an action-oriented one in so far as it views users and learners of a language primarily as "social agents", i.e. members of society who have tasks (not exclusively language-related) to accomplish in a given set of circumstances, in a specific environment and within a particular field of action. While acts of speech occur within language activities, these activities form part of a wider social context, which alone is able to give them their full meaning. We speak of "tasks" in so far as the actions are performed by one or more individuals strategically using their own specific competences to achieve a given result. The action-based approach therefore also takes into account the cognitive, emotional and volitional resources and the full range of abilities specific to and applied by the individual as a social agent.

It is perhaps natural to assume that the components of this description refer only to language use in the "real world" beyond the classroom. But if the target language is to be part of each learner's "everyday lived language", we need to apply those components to the context of learning. The tasks that learners have to accomplish are language learning tasks, the specific environment in which they must perform them is the classroom, their particular field of action is defined by the curriculum, and the social context in which they perform language activities is the community of learners to which they belong. By describing language learners as social agents, the CEFR draws attention to the fact that communicative language use is a matter of making choices, taking decisions, and following through with one or another kind of action. In a classroom intent on extending learners' plurilingual repertoires, learners will exercise agency in relation to their language learning purpose; in other words, they will be taught in such a way that they become autonomous in their learning and use of the target language.

This interpretation of the CEFR's action-oriented approach is not hypothetical but practical, as the following two examples show. In presenting them I have two purposes in mind: (i) to show how languages can be taught and learnt such that from the very beginning they are an integral part of the learners' "everyday lived language", and (ii) to consider how culture enters the picture.

Example 1: Danish teenagers learning English as a foreign language

My first example is widely recognized as the paradigm case of autonomous language learning in a classroom setting. The teacher is Leni Dam, her classroom is in a Danish middle school, her learners are young teenagers (10–15 years old), and the target language is English. Dam's approach is described in detail by Little, Dam and

Legenhausen (2017), who also summarize its theoretical underpinnings and present the results of a longitudinal empirical exploration of the learning outcomes achieved.

Dam first developed her approach in response to the policy of differentiation, adopted by the Danish education authorities in the 1970s. According to this policy, it was the duty of schools to provide a learning environment in which all learners could thrive according to their individual interests and abilities. Learners' interests and abilities are part of their "action knowledge" (Barnes, 1976), the complex of attitudes, beliefs, knowledge and skills that shapes their life outside the classroom. The educational challenge is to use their action knowledge to foster engagement with "school knowledge" (curriculum content), so that what they learn at school gradually becomes a fully integrated part of what they are. When "school knowledge" is a foreign language, the goal from the perspective of the CEFR is to add a new and fully integrated component to the learners' plurilingual repertoires.

The classroom dynamic by which Dam achieved this goal was strongly interactive: collaborative effort was as important as individual effort. From the beginning, learners were expected to make choices and take decisions, which meant that they played initiating as well as responding roles in classroom discourse. The learning activities in which they engaged, individually and collaboratively, always took account of the requirements of the official curriculum, but they also connected with learners' wider interests. In this way learners' cognitive resources, but also their emotional and volitional resources, were fully invested in their language learning. What is more, Dam required them to exercise agency not only in planning, implementing and monitoring successive phases of learning, but in regularly evaluating the learning process and its outcomes. With her support, they did all these things as far as possible in English. Thus, in every lesson the target language was a channel of their agency, and this helped to ensure that it was a fully integrated component of their developing plurilingual repertoires.

From the beginning Dam used English for all classroom communication, and she expected her learners to make every effort to respond to her in English. Especially in the early stages, target language use was strongly dependent on scaffolding and modelling by the teacher, in writing as well as speech. The management of individual and group learning depended on two tools: individual logbooks in which learners kept a record of each lesson, noted new words as they encountered them, wrote the texts they produced individually, and regularly evaluated their learning progress; and posters, written by the teacher in real time in interaction with the class and used for a wide variety of purposes – for example, to compile a list of new vocabulary, record the features of a good conversation, keep track of project work, and evaluate a phase of learning. Learning activities were of two broad kinds: analytic, focused on learning bits of the language (for example, learners made word cards to help them remember words that were particularly important to them), and creative (stories, poems, plays and projects).

Little et al. (2017) provide a wealth of evidence to confirm the success of this approach; here, for reasons of space, I limit myself to just two examples. At the end of

their fourth year of English, when they were fifteen years old, Leni Dam asked the members of one class to write an assessment of their overall progress. Example 1 was written by a boy and Example 2 by a girl; both texts have been transcribed without correction (Dam & Little, 1999):

> **Learner 1**
> Most important is probably the way we have worked. That we were expected to and given the chance to decide ourselves what to do. That we worked independently … And we have learned much more because we have worked with different things. In this way we could help each other because some of us had learned something and others had learned something else. It doesn't mean that we haven't had a teacher to help us. Because we have, and she has helped us. But the day she didn't have the time, we could manage on our own.

> **Learner 2**
> I already make use of the fixed procedures from our diaries when trying to get something done at home. Then I make a list of what to do or remember the following day. That makes things much easier. I have also via English learned to start a conversation with a stranger and ask good questions. And I think that our "together" session has helped me to become better at listening to other people and to be interested in them. I feel that I have learned to believe in myself and to be independent.

These texts are remarkable for their fluency and accuracy – they would not shame 15-year-old native speakers of English. But what they also show is a capacity to use the target language for reflective purposes: surely another criterion for judging whether or not a language is a fully integrated part of a plurilingual repertoire.

How does culture come into the picture? Some of the collaborative projects that Dam's learners undertook required them to engage with English-language culture as "something established" (Byram, 2009:5). But the culture they experienced as "dynamic process" (ibid.) in their language classroom was something they constructed for themselves out of the interaction between their collective "action knowledge" and their collaborative learning effort. After four years of learning English they knew quite a lot about life in Britain and the United States, but they had little if any experience of intercultural encounters of the kind the Council of Europe is interested in promoting. What they did have, however, arising from extensive and sustained collaboration with their peers, was a well-developed capacity to use English as a medium of interaction, joint exploration and negotiation. In due course that capacity was likely to serve them well in their encounters with otherness mediated through English.

Example 2: Plurilingualism and the conversion of linguistic diversity into educational capital

My second example is Scoil Bhríde (Cailíní), a girls' primary school in Blanchardstown, one of Dublin's western suburbs. The school has about 320 pupils; 80% of them come from immigrant families; most of the 80% have little or no English when they start school at the age of four and a half; and between them they have more than 50 home languages. Fuller accounts of the school's highly successful approach to the management and educational exploitation of linguistic diversity are provided by Little and Kirwan (2018a, 2018b) and Little et al. (2017:200–213); a book-length study will be published in 2019 (Little & Kirwan, forthcoming).

Scoil Bhríde's approach to managing extreme linguistic diversity was shaped by two considerations: first, in order to be fully inclusive, the school must find a way of giving immigrant pupils' home languages a role in the life of the school; and, second, whatever measures are adopted, they must provide English-speaking pupils with added value. Three languages were in focus in addition to the immigrant pupils' home languages: English as language of schooling and principal medium of instruction; Irish, the obligatory second language of the curriculum, sometimes used as an alternative medium of classroom communication; and French, taught in Fifth and Sixth Class (the last two primary grades).

Faced with a linguistically diverse pupil cohort, many schools in Ireland and elsewhere insist that all pupils speak only the language of schooling while on the school premises. This policy is usually justified on the apparently reasonable grounds that the more time immigrant pupils spend speaking the language of schooling, the more quickly their proficiency will develop. However, this disregards the fact that whatever language pupils speak at home is central to their identity, the default medium of their discursive thinking, and their most important cognitive tool. Suppressing the use of home languages is, thus, likely to be educationally counter-productive; arguably, it also infringes a basic human right. With these considerations in mind, Scoil Bhríde decided to encourage immigrant pupils to use their home language for whatever purposes seemed appropriate to them, inside as well as outside the classroom.

In the two Infant classes (4–6 years old; equivalent to pre-school in other countries) teachers encourage immigrant pupils to use their home languages to support activities like learning to count. The activity is first performed in English, then in Irish, and then in home languages. Thus, in some classes all pupils learn to count in as many as 15 languages. The games that are an important part of Infant learning are also played multilingually, and again pupils learn fragments of multiple languages (other fragments are picked up playing games in the school yard). Throughout the school teachers present and process curriculum content with frequent reference to home languages – "How do you say that in your language?", "Do you know other languages that have a word for this that sounds like/is very different from the English word?", and so on. This basic pedagogical technique ensures that immigrant pupils' home languages remain

activated in their minds and serves as a constant reminder to native Irish pupils that languages are infinitely diverse.

One further feature of Scoil Bhríde's approach clinches the inclusion of immigrant pupils' home languages in their education: immigrant parents are expected to help their daughters to develop literacy skills in their home language. Texts that are written in the classroom are routinely translated into home languages for homework. To begin with, this involves labels on drawings or a couple of simple sentences copied from the whiteboard; but by the time they are in First Class (6–7 years old) immigrant pupils are beginning to write simple stories in English and their home language. This encourages Irish pupils to write texts in Irish as well as English; then immigrant pupils start writing texts in Irish as well as English and their home language; and that prompts some Irish pupils to learn enough of a third language to keep up with their immigrant peers. In Fifth Class (10–11 years old) French is added to the mix and quickly adopted as another medium in which to write texts. Here are four parallel texts written by a Sixth Class pupil who speaks English with her father and German with her mother:

> Voici Chloë Grace Moretz.
> J'habite á Atlanta, Georgia US. Je la présente, elle est acteur et mannequin. Elle a des yeux bleus et des cheveux blonds. Elle est ni grande, ni petite. Elle est charmante, elle porte un pantalon gris, une chemise bleue, bleu-marine et noire. Elle est intelligente. Elle belle. J'ame son voix, ses vêtements et ses films. Je pense qu'elle est un super actrice.
>
> This is Chloë Grace Moretz.
> Chloë is from Atlanta, Georga, U.S. She has jade coloured eyes and dirty blonde hair. She is of average height. She is wearing grey chinos and pointed heals with a navy, black and blue blouse with a v-neck. I chose Chloë because her choice of style may be preppy but still comfortable.
>
> Seo í Chloë Grace Moretz.
> Is as Atlanta Georgia, usí. Is aisteoir agus mainicíní. Tá gruaig fhíonn uirthí agus tá súile gorm nó glas aicí. Tá gruaig gear aicí. Tá sí idir ard agus beag. Tá sí dathúil. Tá sí ag caitheamh bríste liath agus leine dúghorm, gorm agus dubh. Tá bróga dubh aici freisin. Rhoghnaigh mé Chloë mar tá an éadaí híontach aicí. Is maith liom í ceapaim gur aisteoir í.
>
> Das ist Chloë Grace Moretz.
> Sie ist von Atlanta, Georgia, Amerika. Sie hat blaue Augen und blonde Haare. Sie is von Durchschnittlicher Größe. Sie hat graue Chinos und eine hoher Absätzen auf eine marine schwarz und blaue Bluse. Ich entschiede mich für Chloë, weil ihre Wahl der Stil ist vieleicht formal, aber immer Komfortabel.

Only the English text is without blemish, but all four texts were written spontaneously and without hesitation. Like the English texts produced by the two Danish learners, they are evidence of a fully integrated plurilingual repertoire: English, Irish, German and French are this girl's "everyday lived language(s)".

The relation between language and culture here is much the same as in the Danish example. By encouraging immigrant pupils to use their home languages, Scoil Bhríde engages their action knowledge with the educational process. As pupils learn and use English, Irish and French and develop literacy skills in their home language, they encounter a vast array of cultural fragments as "something established". Some of those fragments are already familiar to them from their lives outside school; others are new and perhaps strange. Culture as dynamic process is something they and their teachers create for themselves in the constantly shifting multilingual interactions of the classroom. Scoil Bhríde's approach to language education enables pupils to develop unusually sophisticated levels of language awareness, and their critical interest in language carries over into other areas of the curriculum. Combined with their plurilingual fluency, this gives them the potential to engage successfully in intercultural encounters later in life.

5. Conclusion

In this article I have explored the contradictions in which the CEFR entangles its concept of plurilingualism, explained why pluriculturalism is an impossible educational goal, and described two language learning environments in which learners undoubtedly develop plurilingual repertoires, but without explicitly engaging with cultural otherness. In formal language learning contexts, plurilingualism without interculturality seems to me an inevitable consequence of the fact that plurilingualism is by definition always and necessarily the product of the language user/learner's here-and-now.

As promoted by the Council of Europe, plurilingual and intercultural education is seriously under-determined: none of the organization's guides and discussion papers descend from general abstraction to consider in detail the practicalities of pedagogical implementation. If we treat plurilingual and intercultural education as a working hypothesis rather than a new educational orthodoxy, however, we can subject it to empirical investigation and thereby gain greater clarity and perhaps make progress. To this end, we need more research that explores the relation between plurilingual language learning and intercultural education *in practice*; research, moreover, that takes account of publications critical of the Council of Europe's stance (e.g. Dervin, 2016) and pays attention to the argument of Wood (2003:21) that:

> we are drunk with the idea that every difference of ethnic custom, every foreign or regional accent, every traditional recipe, and every in-group attitude betokens a distinct world view.

Let me end on a personal note. I first met Sauli Takala in 1981 at the Applied Linguistics World Congress in Lund, where we both contributed to a Council of Europe symposium organized by John Trim. Subsequently we met at regular intervals, most often at Council of Europe events in Strasbourg; we shared a strong but by no means uncritical

commitment to the CEFR. The arguments presented in this article were taking shape when I heard of Sauli's sudden death. One of my first thoughts was that I would not now be able to share them with him. I like to think, however, that if I had been able to rehearse them over dinner and a few beers, he would have found them interesting, worth discussing and perhaps worth criticizing.

References

Barnes, D. (1976). *From Communication to Curriculum*. Harmondsworth: Penguin.

Byram, M. (2009). Multicultural societies, pluricultural people and the project of intercultural education. Strasbourg: Council of Europe. Available: https://rm.coe.int/ CoERMPublic CommonSearchServices/DisplayDCTMContent?documentId=09000016805a223c

Cook, V. J. (1991). The poverty-of-the-stimulus argument and multi-competence. *Second Language Research* 7(2), 103–117.

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

Dam, L. & Little, D. (1999). Autonomy in foreign language learning: From classroom practice to generalizable theory. In A. W. Barfield, R. Betts, J. Cunningham, N. Dunn, H. Katsura, K. Kobayashi, N. Padden, N. Parry & M. Watanabe (Eds), *Focus on the Classroom: Interpretations*, Proceedings of the 24th JALT International Conference, 127–36. Tokyo: The Japan Association for Language Teaching.

Dervin, F. (2016). *Interculturality in Education: A Theoretical and Methodological Toolbox*. London: Palgrave Macmillan.

European Commission (2012). *First European Survey on Language Competences: Executive summary.* Brussels: EU Commission. Available: http://www.surveylang.org/media/ ExecutivesummaryoftheESLC_210612.pdf

García, O. (2017). Problematizing linguistic integration of adult migrants: The role of translanguaging and language teachers. In J.-C. Beacco, H.-J. Krumm & D. Little (eds), *The Linguistic Integration of Adult Migrants / L'Intégration Linguistique des Migrants Adultes. Some Lessons from Research/Les Enseignements de la Recherche*, 11–26. Berlin: de Gruyter.

Little, D., Dam, L. & Legenhausen, L. (2017). *Language Learner Autonomy: Theory, Practice and Research*. Bristol: Multilingual Matters.

Little, D. & Kirwan, D. (2018a). Translanguaging as a key to educational success: The experience of one Irish primary school. In K. Maryns, S. Slembrouck, S. Sierens, P. Van Avermaet, K. Van Gorp (Eds), *The Multilingual Edge of Education*, 313–339. Basingstoke, UK: Palgrave Macmillan.

Little, D. & Kirwan, D. (2018b). From plurilingual repertoires to language awareness: Developing primary pupils' proficiency in the language of schooling. In C. Frijns, K. Van Gorp, C. Hélot & S. Sierens (Eds), *Language Awareness in Multilingual Classrooms in Europe*, 169–205. Berlin: Mouton de Gruyter.

Little, D. & Kirwan, D. (forthcoming). *Engaging with Linguistic Diversity: A Study of Educational Inclusion in an Irish Primary School*. London: Bloomsbury Academic.

Wood, P. (2003). *Diversity: The Invention of a Concept*. San Francisco: Encounter Books.

”In one sentence there can easily be three different languages”. A glimpse into the use of languages among immersion students[37]

Karita Mård-Miettinen
University of Jyväskylä

Siv Björklund
Åbo Akademi University

1. Introduction and aim of the study

In an institutional context, multilingual development has traditionally been studied from the point of view of a linguistic minority. The studies have, for example, focused on the extent to which it is possible for students with immigrant background or students from a linguistic minority to use their own language in day care or school in a majority language (e.g. Martin-Jones & Martin, 2017). The development of multilingual skills among a student that belongs to a linguistic majority is not obvious in the same way as it is for a student belonging to a linguistic minority, because the linguistic majority student may not have a natural need to use other languages than his first language outside school. This is the situation for Swedish immersion students in Finland, for example, since they are nearly without exception linguistic majority students who are supposed to develop multilingual skills in school. One of the downsides of immersion education is that the use of the immersion language might be limited only to the school context and does not become a part of the everyday life of the immersion students outside the school (Johnson & Swain, 1997). Immersion research gives, thus, a novel point of view to research on multilingualism and a possibility to investigate the roles of languages taught in school in young people’s lives.

[37] The article is originally published in Finnish in Mård-Miettinen, K., & Björklund, S. (2018). "Yhes lausees saattaa olla ihan helposti kolmee eri kieltä": kurkistus kielikylpyoppilaiden kielimaailmaan. In L. Nieminen, A. Yliherva, J. Alian, & S. Stolt (Eds.), *Monimuotoinen monikielisyys: Puheen ja kielen tutkimuksen päivät Helsingissä 5.-.6.4.2018* (pp. 80-91). Puheen ja kielen tutkimuksen yhdistyksen julkaisuja, 50. Puheen ja kielen tutkimuksen yhdistys. Translation: MA Sannina Sjöberg.

The aim of the study reported in this article is to deepen the understanding of multilingualism among immersion students and to investigate how the students express their use of languages visually and in an elicitation interview based on the visual data. Visual methods are based on drawings, commercials or films that are produced by either the researcher or the subject of the study or that are naturally occurring visual products (Heath, Brooks, Cleaver & Ireland, 2009). Visual methods have been employed in the field of language research for only a relatively short time, but they have a long tradition in social sciences in researching the social worlds of everyday life (Pitkänen-Huhta & Pietikäinen, 2017; Rose, 2016). In recent years, visual methods have been used more frequently also in ethnographic research concerning language learning and language use. The study reported in this article represents this type of research.

2. The many dimensions of multilingualism

Research in multilingualism is a fast-growing research field, and one of its most important results is a more pragmatic viewpoint to the different conceptions of language and its situational use (e.g. Dufva & Pietikäinen, 2009). The fact that multilingualism is so multifaceted makes it a challenging research object. When doing research on multilingualism, the focus is on the dialogic relation between different languages, whereas research on bilingualism is often centered on a comparison of the mastery in the two languages involved. As many researchers of multilingualism, we do not consider it appropriate to have monolingualism as a starting point, but rather a multilingual language user, the whole language repertoire of the individual as well as the context (cf. Cenoz, 2013). We thus assume that the individual's own understanding of his/her languages and the use of them reflects the surrounding linguistic landscape, contexts where the person operates and personal contacts that s/he has.

For institutional purposes, languages are often categorised. In immersion, as in the school context generally, languages have symbols that describe them (mother tongue, immersion language, foreign language, etc.) and they have their own places in the students´ timetables. From the point of view of the language user, these categories and symbols do not have a meaning, and Pennycook (2006) has stated that languages can be at the same time static and bounded, and dynamic and mixed. In immersion education, the aim is functional bilingualism and multilingualism, even though there are features in the immersion programme that highlight parallel monolingualism (the concept of parallel monolingualism, e.g. Heller, 1999). When defining immersion education, one of the core features is that the teaching of each subject only is in one language at a time, as well as each teacher acting as a monolingual language model (e.g. Bergroth, 2015). In the new national core curriculum for basic education in Finland, it is emphasised that also the teaching material in immersion should be in the same language as the teaching (Finnish National Board of Education, 2014). This principle is created to support the development of the new language, the immersion

language, in particular, into a strong language to be used for learning subject matter content. It also creates as diverse and natural situations as possible for using the immersion language. (Bergroth, 2015) The principle of parallel monolingualism is, however, not extended to immersion students. Rather, they are allowed to use all their linguistic resources to support their learning. In research literature, it is highlighted that especially during the lessons taught in the immersion language, students should be encouraged to use the immersion language for it to develop into as strong a language for learning as possible (e.g. Bergroth, 2015).

Previous research on multilingualism among immersion students in Finland has been based on survey and interview data. Immersion students in grades 4–9 (385 students) in different parts of Finland answered questions about language learning and studies and language knowledge and use. The study showed that many immersion students study optional languages in school (Björklund & Mård-Miettinen, 2011a and 2011b). The immersion students also reported that they knew all the languages that they studied in school, and most of them considered themselves as multilinguals (Björklund & Mård-Miettinen, 2011b; Björklund, Pakarinen & Mård-Miettinen, 2015).

Survey-based research on beliefs give, however, only limited access to students' language use outside school (cf. Pitkänen-Huhta & Pietikäinen, 2017). Survey and interview data collected in a school context may distort the results since the students may consider only languages taught in school and not all the languages they know, and those they use outside school. This has been the case in previous immersion research based on survey and interview data: immersion students tended to report their language use only for those languages they studied in school (Björklund & Mård-Miettinen, 2011b). In this article, we approach the topic via visual data that is a novel method in studying the language use of immersion students outside the school context.

3. Participants, data and methodology

The participants in our study were ten students in grades 5 and 8 in basic education, three girls and two boys from each grade level. Their ages were 11-12 years (grade 5) and 14-15 years (grade 8) at the time of the study. The participants had attended early a total immersion programme in Swedish in southern Finland since the age of four years. Early total immersion is a programme carried out in Finnish early childhood education (ages between 3 and 5 years), preschool education (for 6 year-olds) and basic education (grades 1-9, ages 7 to 16 years), where children from Finnish speaking families get the opportunity to study several languages and obtain functional multilingualism (e.g. Björklund & Mård-Miettinen, 2011). In the immersion programme the participants in the study attended, the language of instruction in all activities both in early childhood education and in preschool was the immersion language, Swedish. In the initial grades of the basic education, most of the subjects were taught through Swedish, and some subjects were taught in Finnish. The relationship between the teaching in Swedish and in Finnish changed during the school path so that the subjects taught in Finnish were about 50 % by grade 5. Further, the

students were taught several foreign languages since the initial grades in basic education. All the participants had started studying English at grade 3, and all except one student had started studying Spanish or German in grade 4. Two of the students who were in grade 8 at the time of the study, had just begun studying Spanish or German as an elective. At the time of the study, the students were studying half of their school subjects through Swedish and half of them in Finnish. All students were studying English and either Spanish or German, and one student in grade 8 was studying three foreign languages in addition to Finnish and Swedish (English, Spanish and German). In a background survey, two students reported knowing also French and Estonian, which they did not study at school.

The data were collected using visual methods[38]. The students were instructed to take photographs of their use of different languages with their mobile phones. Pitkänen-Huhta and Pietikäinen (2017) emphasise that through visual data you can make visible language experiences and practices without the need to use restricting classifications of languages or language skills. The data in our study were produced by the participants, the immersion students. According to Heath et al. (2009), using data produced by participants makes the participants active agents and opens access to more private spaces than other methods.

Our data collection started with a short background survey to the participants in which they listed languages they studied and knew. After that, they received instructions on how to take the photographs. The students were asked to take 2–3 photographs a day during a week, in their free time and on school breaks illustrating typical situations where they used their different languages. The students sent their photographs to the researchers by e-mail and added a short comment to each photograph. The number of photographs sent by each student varied between two and eleven pictures, and our research data consist of 71 photographs in total. Two weeks after the taking of photographs, the students participated in a 15 minute structured individual photo elicitation interview, where they were asked to talk about the photographs they took from the point of view of language use. The photo elicitation interview was expected to give a deeper understanding of how the photographs were related to the students' use of different languages. Heath et al. (2009) emphasise that the visualisation itself does not necessarily reveal the thoughts of the participant, and therefore it is important for the researcher to have the interpretation and the point of view of the photographer to support the analysis.

In our analysis, we examined the visual data in conjunction with the elicitation interview data. In the first phase, we applied domain analysis (Fishman, 1972) and identified the languages that appeared in the students´ photographs and interviews, and spaces for language use (places, situations, actions and persons) that the students related to these languages. Through this analysis, we aim to understand which languages belong to the students' language repertoire and how they appear in their

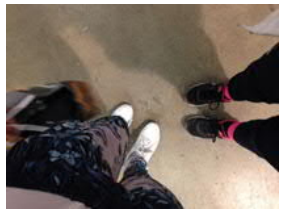
[38] The research reported in this article belongs to a larger research project on language practices, multilingual identity and language ideology in immersion education, financed by the Society of Swedish Literature in Finland (SLS, http://www.sls.fi/en).

lives. In the second phase, we examined in more detail how the students expressed their language use. The analysis focused on discourses about language separation and dynamic language use. Considering that immersion education is strongly based on parallel monolingualism, we found it interesting to study how this separation of languages in immersion education possibly reflects in the students´ discourses of their language use.

4. Results

4.1 Languages and spaces for language use

Our data indicate that the immersion students in our study live a rather multilingual life. Those students that had taken pictures at school, had photographed their language use at breaks and at an exhibition visit from school. In these domains, the students used two or three languages with classmates, friends and exhibitors. Some students said they were using three languages in these domains (Finnish, Swedish and English) and some said they used two languages (Finnish and English):



Kyl siel välillä tulee aina englanninkielisii sanojaki mukaa sinne keskusteluun mut yleensä suomeks aina puhutaan [kavereiden kanssa välitunneilla]. (Paula, 8. lk)
Sometimes English words appear in the discussion but usually we always speak Finnish [with friends on breaks]. (Paula, grade 8)

Several students took pictures of their language use with friends also outside school in a similar way like Paula in the above example from a school break. When talking with friends who know Swedish, the students said they used Swedish as well as English words.

Altogether, more languages were related to spaces outside school than to spaces in school in our data. In addition to the languages studied at school (Finnish, Swedish, English, Spanish and German), languages related to the students' leisure time were Chinese, Estonian, French, Japanese, and Norwegian. All students used Finnish, Swedish and English in their leisure time and most of them in several places, situations and activities, and regularly. Especially the use of English was strongly related to the daily watching of programmes, even though some students watched programmes also in other languages, as reported by Arttu:

Toi on yks ohjelma englanniks mitä mä katon melkein joka aamu...viikossa mä katon jakson [virolaista komediaohjelmaa] joka kestää tunnin. (Arttu, 5. lk)
That is one programme in English that I watch almost every morning…in a week I watch an episode of [an Estonian sitcom] that takes one hour. (Arttu, grade 5)

Swedish and English were used regularly also for playing games and reading books. Spanish and German (languages taught at school) were also used regularly, but mainly when doing homework. Other languages that appeared in our data were languages that a student had noticed in his/her environment, like Mea when playing a board game at home with her family:

Me pelattiin illalla sellasta peliä sit pelin reunass luki melkein kaikilla kielillä että opi ja leiki ja pelaa ja tutki...siel oli englantia, saksaa ja suomea ja ruotsia ja norjaa ja japaniiki tais olla. (Mea, 5. lk)
We played a game in the evening and it read on the game cover in almost every language learn and play and play and explore… there was English, German and Finnish and Swedish and Norwegian and there may have been Japanese, too. (Mea, grade 5)

The students reported actively using other languages than the languages studied at school as reported by Anton:

> *Yhden norjalaisen kaa mä ain välillä puhun ... me puhutaan välil sillee et se puhuu norjaa ja mä puhun ruotsii ja sit me ymmärretään toisiamme et ne on niin samankaltasii kielii kuitenkin ja muuten englanniks jos ei ymmärrä.* (Anton, 8. lk)
> *I sometimes talk to a Norwegian… sometimes s/he speaks Norwegian and I speak Swedish and we understand each other because the languages are so alike and sometimes in English if we don´t understand each other.* (Anton, grade 8)

Several students in our data reported that they do as Anton does, i.e. they use their Swedish skills to understand oral or written Norwegian.

Altogether, the students described using their different languages in many places, situations, activities and with many different people. The languages were used at home and in hobbies, with family members, relatives, friends, hobby mates, game mates as well as pets. The activities in which the students used their languages, were doing homework, reading books and magazines, watching movies and TV, listening to music, searching on the internet and watching Youtube clips, playing games (computer and board games) and using social media. Some students had hobbies where several languages were used:

Meiän joukkuessa on yks englanninkielinen joka puhuu vähän suomee mut sen kans puhutaan englantii sit siel on yks norjalainen jonka kanssa puhutaan ruotsia tai englantii. (Arttu, 5. lk)
In our team there is one English-speaking person who speaks a little Finnish but with him we speak English, then there is one Norwegian with whom we speak Swedish or English. (Arttu, grade 5)

Tanssitermeiki on itse asias eri kielillä et siin nyt tulee ranskaaki ja kaikkee mut kyl tanssitermitki on englanniks. (Siru, 8. lk)
Dance terms are in fact in different languages so there is also French and everything but the dance terms, too, are in English. (Siru, grade 8)

4.2 Discourses of language separation

There are numerous mentions in our data that emphasise that a certain language is used in a described activity or situation:

Mä pelaan tietokonepelejä kansainvälisesti et kaikkien kaa. Yleensä mä kommunikoin englanniks ja sit välillä ruotsiks kans. Se riippuu kenen kaa mä pelaan just. (Anton, 8. lk)
I play computer games internationally so with everyone. Usually I communicate in English and sometimes in Swedish, too. It depends on who I´m playing with at the moment. (Anton, grade 8)

Anton describes above that he uses English and Swedish while playing games, but he mentions separating the languages according to person suggesting that he communicates monolingually when playing games. Anton as well as other students in our data separate languages according to situation, i.e. the activity in focus is not related to only one specific language in the student´s life, but the same activity can be completed in several languages. This concerns, except playing, also e.g. reading and watching programmes. The unifying factor is that the activities are presented as monolingual activities in the interview.

Our data even includes descriptions of the students using two different languages parallel to doing different things in each language. A frequent activity related to the parallel use of two languages is watching a movie or a TV programme in English with Finnish subtitles:

Me käytii kattoo leffa joka oli englanniks niin sitte tuli eri kielii käytettyy sillee et mä kuulin englantii ja sitten luin sen suomeks sen tekstin. (Minea, 8. lk)
We went to see an English speaking film so then different languages were used because I listened to English and read the subtitles in Finnish. (Minea, grade 8)

Some of the students commented watching movies in English using subtitles, as Minea, but some of them also watched movies monolingually without subtitles. Furthermore, some students reported having the subtitles on but barely looking at them, like Rickard:

> *Katon teeveetä koulun jälkeen se on englanniks ja mä en yleensä jos mä katon jotain englanniks niin mun ei yleensä hirveesti tartte kattoo niitä tekstityksiä.* (Rickard, 5. lk)
> *I watch the TV after school, it´s in English and usually when I watch something in English I don´t usually need to read the subtitles much.* (Rickard, grade 5)

In addition to the students using different languages simultaneously for listening and reading, two different languages were also produced parallel for speaking and writing:

> *Mä puhuin suomee sillee [kaverille skypessa] ja samalla mä kirjoitin muille englanniks siinä pelissä.* (Anton, 8. lk)
> *I talked like in Finnish [to a friend on Skype] and at the same time I wrote to others in English in the game.* (Anton, grade 8)

4.3 Discourses of dynamic language use

Visual data also revealed situations of dynamic language use. Some of the immersion students reported mixing different languages particularly when talking to their friends, as in the example with Siru:

> *Ku jutellaan vaan ihan vaan jostain random asioista niin sitte saattaa sanoo jotain sanoja englanniks ... en mä ees huomaa et mitä kielii mä puhun varsinki jos näitten tiettyjen kavereitten kaa et jos ne osaa mun kaa suomee, ruotsii ja englantii niin saattaa kaikki kielet mennä vaan yhtäkkii sekaisin ... yhes lausees saattaa ihan helposti olla kolmee eri kieltä.* (Siru, 8. lk)
> *When we talk just about some random things then you might say some words in English… I don´t even notice what languages I speak especially with these specific friends like if they know with me Finnish, Swedish and English then suddenly all languages might get mixed up… in one sentence there can easily be three different languages.* (Siru, grade 8)

Siru emphasises above that the dynamic use of several languages requires that the conversation partners know the languages. Like Siru, several other students mentioned that especially English words appear when they speak in Finnish with friends or when playing computer games. In Siru´s comment above a subconscious simultaneous use of several languages is emphasised. It is notable, that she even uses the English word ‘random’ in her otherwise Finnish language response in the comment above. However, a majority of the students mentioned consciously using elements from another language when speaking in one language in some situations, as Ada when packing her backpack with her mother:

> *Siinä mä pakkasin mun reppua ja samalla puhuin äidille suomeks ja sitte jotain mä yritin saksaks sanoo ku mulla oli maanantaina saksan tunti.* (Ada, 5. lk)

> *There I was packing my backpack and at the same time talking in Finnish to my mum and then I tried to say something in German because I had a German class on Monday.* (Ada, grade 5)

When using other languages than Finnish, the students often mentioned that Finnish is needed to help express things:

> *Mä olin siinä tekemässä historian läksyjä ruotsiksi ja sit suomeksi vähän kun en osannut ruotsiksi sanoo.* (Aku, 8. lk)
> *There I was doing my history homework in Swedish and then a little bit in Finnish when I didn´t know how to say it in Swedish.* (Aku, grade 8)

Particularly when it comes to doing homework, the students often report using Finnish as a backup language. One student also emphasised that dynamic language use was related especially to situations where other languages than Finnish were used:

> *En mä yleensä puhu muit kielii [kuin suomea] sillee kokonaan et jotain sanoja sit.* (Minea, 8. lk)
> *I usually don´t speak other languages [than Finnish] like completely so just some words.* (Minea, grade 8)

5. Discussion and conclusions

The life of the immersion students in our study was multilingual; reflecting at least partly their immersion experience. Photographs taken during school breaks and trips highlighted the languages studied for the longest time in school, that is the mother tongue Finnish, the immersion language Swedish and their first foreign language, English. The students used these languages with their classmates and friends. None of the students described their language use with teachers or other members of school staff outside the classroom, even though it can be assumed that these situations appear in school.

The questionnaire on the students´ language skills and use of languages gave a similar picture of the students' language repertoire as in previous survey and interview studies; basically only the languages studied at school were mentioned (cf. Björklund & Mård-Miettinen, 2011a and 2011b). The visual data in this study showed that the contexts for language use in the students' leisure time were diverse and connected even to other languages than the languages studied at school.

Previous immersion research (e.g. Björklund & Mård-Miettinen, 2011a) has shown that English emerges next to the immersion language Swedish as an important language to the students at a very early stage. This showed also in this study. The immersion students in our study used English for many purposes in their leisure time and often daily. This confirms results in Finnish school-based research that have shown

that the use of English is substantial in the students' leisure time (e.g. Pitkänen-Huhta & Nikula, 2008).

The immersion students, however, also used Swedish when reading, during their hobbies, on social media and with friends, family and relatives, i.e. Swedish has become a part of many immersion students' everyday life also outside school. Swedish has also brought Norwegian into many immersion students' everyday life. The students reported reading and listening to Norwegian using Swedish themselves in these situations, and when necessary using English as support. Additionally, some students in our study described situations where they used languages they do not study at school, such as French, Estonian or Chinese. Some students also described situations where they had noticed the presence of certain languages in their environment, that they did not know themselves (for example Japanese). The foreign languages studied at school, German and Spanish, also appeared regularly in the immersion students' everyday life, but mainly in connection to doing homework. These languages belong to the life of the students at least for as long as they study them at school. These languages were also used somewhat on holidays.

The core immersion feature of parallel monolingualism does not seem to have had a remarkable effect on the immersion students' way of using their languages in their leisure time since the students' descriptions included discourses of both language separation and dynamic language use. Concerning language separation, the students talked about using one language at a time, so that a specific language was used with a specific person or in a specific situation or activity. The same activity was often reported as carried out in several languages. Additionally, the students gave examples of the parallel use of two languages, so that they simultaneously listened in one language and read in another language or spoke in one language and wrote in another language.

The immersion students mentioned the dynamic use of different languages mainly when they described talking with their friends. In these situations, they explained English and/or Swedish words appearing in their Finnish speech. The students also reported needing Finnish as support when doing their homework in Swedish.

Employing visual methods when collecting data turned out to be successful and the students were highly motivated to take photographs as well as to describe their language use even outside the situations in the pictures. A larger data would, however, be needed for further conclusions. Our data still gives a glimpse of the diverse and multilingual life of some immersion students in the way they have selected to picture it: *That I can function in many languages completely normally.* (*Et mä pystyn toimimaan monella kielellä ihan normaalisti,* Minea, grade 8)

References

Bergroth, M. (2015). *Kotimaisten kielten kielikylpy*. Vaasan yliopiston julkaisuja. Selvityksiä ja raportteja 202. Vaasa: University of Vaasa. Available at: https://www.univaasa.fi/materiaali/pdf/isbn_978-952-476-617-3.pdf.

Björklund, S. & K. Mård-Miettinen (2011a). Integration of multiple languages in immersion: Swedish immersion in Finland. In: D. J. Tedick, D. Christian & T. Williams Fortune (eds.), *Immersion education: Practices, policies, possibilities*, (p. 13–35). Multilingual Matters.

Björklund, S. & K. Mård-Miettinen (2011b). Kielikylpylasten ja -nuorten monikielinen toimijuus. In: N. Mäntylä (eds.), *Lapset ja nuoret yhteiskunnan toimijoina* (p. 154–169). Vaasan yliopiston julkaisuja. Tutkimuksia 297. Vaasa: University of Vaasa. Available at: https://www.univaasa.fi/materiaali/pdf/isbn_978-952-476-379-0.pdf.

Björklund, S., S. Pakarinen & K. Mård-Miettinen. 2015. Är jag flerspråkig? Språkbadselevers uppfattning om sin flerspråkighet. In: J. Kalliokoski, K. Mård-Miettinen & T. Nikula (eds.), *Kieli koulutuksen resurssina: vieraalla ja toiselle kielellä oppimisen ja opetuksen näkökulmia* (p. 153–167). AFinLA-e. Soveltavan kielitieteen tutkimuksia 2015/n:o 8. Jyväskylä: Jyväskylän yliopisto. Available at: http://ojs.tsv.fi/index.php/afinla/article/ view/ 53777/16874.

Cenoz, J. (2013). The influence of bilingualism on third language acquisition: Focus on multilingualism. *Language teaching* 46:1, 71–86.

Dufva, H & S. Pietikäinen (2009). Moni-ilmeinen monikielisyys. *Puhe ja kieli* 29:1, 1–14.

Fishman, J. A. (1972). *Language in sociocultural change*. California: Stanford University Press.

Heath, S., R. Brooks, E. Cleaver & E. Ireland (2009*). Researching young people's lives*. Sage Publications Ltd.

Heller, M. (1999). *Linguistic Minorities and Modernity: A Sociolinguistic Ethnography*. London: Longman.

Johnson, R. K. & Swain, M. (1997). Immersion education: a category within bilingual education. In: R. K. Johnson & M. Swain (eds.), *Immersion education. International perspectives* (p. 1–16). Cambridge: Cambridge University Press.

Martin-Jones, M. & Martin, D. (2017). Introduction. In: M. Martin-Jones & D. Martin (eds.), *Researching multilingualism. Critical and ethnographic perspectives* (p. 1–27). Oxon: Routledge.

Opetushallitus (2014). *Perusopetuksen opetussuunnitelman perusteet*. Helsinki: Opetushallitus. Available at: http://www.oph.fi/saadokset_ja_ohjeet/ opetussuunnitelmien_ja_tutkintojen_perusteet/perusopetus.

Pennycook, A. (2006). *Global Englishes and transcultural flows*. Abingdon: Routledge.

Pitkänen-Huhta, A. & Nikula, T. (2008). Using photographs to access stories of learning English. In: P. Kalaja V. Menezes & A.M.F. Barcelos (eds.), *Narratives of Learning and Teaching EFL* (p. 171–185). Basingstoke: Palgrave Macmillan.

Pitkänen-Huhta, A. & Pietikäinen, S. (2017). Visual methods in researching language practices and language learning: Looking at, seeing, and designing Language. In: K. King, Y.-J. Lai & S. May (eds.), *Research methods in language and education, Encyclopedia of language and education* (p. 393–405). Basel: Springer International Publishing.

Rose, G. (2016). *Visual methodologies. An introduction to researching with visual datas*. 4 th edition. London: Sage Publications Ltd.

Redefining specific purpose tests

Barry O'Sullivan
British Council

1. A brief historical introduction

Though the teaching and testing of general language proficiency has been around for many years (see Weir & Milanovic, 2003), interest in language for specific purposes has a far shorter history. According to Swales (1984:11) it emerged with Barber's (1962) *Some Measurable Characteristics of Modern Scientific Prose*, though there has long been an awareness of the use of language for specific purposes (LSP), as we are reminded by Schröder (1981:43) who reports on the language studies in Britain of young apprentices from Germany in the 16th century.

Much early work in the area was motivated by research which focused on: (i) the identification of unique instances of language use in specific contexts (Swales, 1971; Lackstrom, Selinker & Trimble, 1973; Johns, 1980; Hüllen, 1981a, 1981b; Selinker & Douglas, 1985, to list but a few); (ii) the issue of authenticity in the use of materials for teaching (e.g. Carver, 1983); and (iii) the central place of needs analysis in identifying the specific language needs of learners in given contexts (LCCIEB, 1972; Alwright & Alwright, 1977; Brindley, 1984; Gledhill, 2000; Hawkey, 1978; Hutchinson & Walters, 1987; Kennedy & Bolitho, 1984; Robinson, 1980, 1985; Thurstun & Candlin, 1998; West 1994).

In the case of the testing of language for business purposes, the first test to emerge was the Test of English of International Communication (TOEIC). It was developed by the Educational Testing Service (ETS) in the United States of America and introduced in 1979. The test, originally devised for the Japanese market, was based on psychometric-structuralist theory (Spolsky, 1995) and represents one of the few remaining examples of a purely multiple-choice format, standardised, international language test.

While the TOEIC looked backwards for its theoretical underpinning, other tests of business language, particularly those developed in the United Kingdom, were beginning to look to a more communicative model. Theorists in the area of communicative competence, particularly Hymes (1972), Canale and Swain (1980) and practitioners like Munby (1978) had a profound influence on the practice of language teaching and testing. One major influence was the facilitation of a movement away from the psychometric-structuralist methodology, based on the teaching and testing of discreet aspects of language, to the psycholinguistic-sociolinguistic era, where language teaching and testing were seen from a holistic or integrated perspective. The

shift in emphasis in language teaching from language *knowledge* to language *use* paved the way for a testing methodology which reflected the same ideas.

In the mid-1980s, the move to the testing of language for business purposes in the United Kingdom began in earnest with the development by the Royal Society of Arts (RSA) of the *Certificate in English as a Foreign Languages for Secretaries* (CEFLS) – which was later administered as the Certificate in English for International Business and Trade (CEIBT), and a corresponding moves by both the London Chamber of Commerce and Industry Examinations Board (LCCIEB) and Pitman to create their own language tests with a business focus. By the early 1990s new examinations, such as the Business English Certificates (BEC) were developed by the University of Cambridge Local Examinations Syndicate (UCLES). In more recent years a number of tests of other languages for business emerged. These included JETRO (Japanese), Test de français international (TFI) from the makers of TOEIC, the Certificate in Italian for Commerce (CIC) and the tests in the BULATS series (French, German and Spanish in addition to the English version). See O'Sullivan (2006) for a full review of these tests.

There is clearly a growing interest in the area of testing language for business purposes, particularly with the internationalisation of business and the need for employees to interact in more than just a single language.

2. Theoretical Perspectives

Douglas (2000) argues that a theoretical framework can be built around two principal theoretical foundations. The first of these theoretical foundations is based on the assumption that language performance varies with the context of that performance. This assumption is supported by a well established literature in the area of sociolinguistics in addition to research in the areas of second language acquisition (Dickerson 1975; Ellis 1989; Schmidt 1980; Smith 1989; Tarone 1985, 1988) and language testing (Berry, 1996, 1997, 2004; Brown, 1995, 1998; Brown & Lumley, 1997; O'Sullivan 1995, 2000a, 2000b, 2002; Porter 1991a, 1991b; Porter & Shen, 1991). This fits well with the growing interest in a socio-cognitive approach to language test development where performance conditions are seen to have a symbiotic relationship with the cognitive processing involved in task completion (Weir, 2005; O'Sullivan & Weir, 2011; O'Sullivan, 2011, 2014, 2016).

In the case of the second theoretical foundation, Douglas (2000???) sees specific purpose language tests (SPLTs) as being 'precise' in that they will have lexical, semantic, syntactic and phonological characteristics that distinguish them from the language of more 'general purpose' contexts. This aspect of Douglas' position is also supported by a significant literature, most notably in the area of corpus-based studies of language in specific contexts (Beeching, 1997; Biber et al., 1998; Dudley-Evans & St John, 1996; Gledhill, 2000; Thurstun & Candlin, 1998. Hyland, 2007, 2009, 2012, 2015).

When it comes to an actual definition of specific purpose tests, Douglas (2000:19) places these two foundations within a single over-riding concept, that of authenticity, defining a test of specific purposes as:

> One in which test content and methods are derived from an analysis of a specific purpose target language use situation, so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test taker's language ability and specific purpose content knowledge, on one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain.

This definition highlights the core element of Douglas' view of LSP tests; that of *authenticity*. Douglas does not see this as being a simple matter of replicating specific purpose tasks in a testing context, but of addressing authenticity from two perspectives. The first perspective is that of *situational* authenticity, where LSP test tasks are seen as being 'authentic' in that they are derived from an analysis of the language use domain with which they are associated. The second perspective is *interactional* authenticity, which relates to the actual processing that takes place in task performance, what Weir (2005) refers to as theory-based validity.

This definition has not remained unquestioned. In fact, Douglas (2001) acknowledges that there are a number of issues left unanswered by his definition, an argument also made by Elder (2001). This criticism focuses on what Elder (2001) sees as the three principal problematic areas identified in the work of Douglas, namely, the distinguishability of distinct 'specific purpose' contexts; authenticity; and the impact (and interaction) of non-language factors.In terms of the latter point, it can be argued that a test of language for a specific purpose should not even try to avoid the background knowledge issue, as it is this that defines the test. How we deal with the situation will depend on the degree of specificity of the test and the inferences we intend to draw from performance on the test. Turning, therefore, to the remaining criticisms of an ESP approach to testing, we can see that there are basically two questions that should be addressed, these are:

- Distinguishing LSP from general language – is it possible and/or feasible?
- Authenticity – can LSP tests be made both situationally and interactionally authentic?

3. Distinguishing LSP from General English

There is a considerable body of work over the last 30 years which has quite clearly demonstrated the distinguishability of language use in specific contexts. We can point to the work on the definition of language needs and usage in specific contexts of needs analysis researchers and theorists. Among the influential early work were studies undertaken by Hawkey (1978) who demonstrated how needs analysis can lead to a

specific purpose curriculum, as well as Alwright and Alwright's (1977) practical advice on an approach to the teaching of medical English.

In the area of testing language for specific purposes, perhaps the most important undertaking was that of the London Chamber of Commerce and Industry Examinations Board (LCCIEB) in 1972. The LCCIEB had been providing business-related qualifications around the world for almost 100 years when, in 1972, its Language Section undertook a major analysis of 'foreign' language use involving over 11,500 employees of almost 600 international firms. This analysis -together with the replications undertaken in the Federal Republic of Germany, France, Greece and Spain between 1982 and 1985 - was to prove influential in the development of teaching and testing practice in the UK during the 1970s and 1980s.

Studies such as those carried out by Alderson and Urquhart (1984, 1985, 1988) and Steffensen and Joag-dev (1984) suggested that background knowledge is a significant factor in specific purpose language testing, a point that was also made by Clapham (1996) with reference to highly specific tests.

The implication of the work referred to earlier in the paper, when seen in light of this small but important body of work, is that there is a clearly definable language of business (and of other areas of specific interest such as science, technology etc.) and that where tests are devised with a deliberately high level of specificity towards an explicit area, then candidates whose background is grounded in that area can be expected to outperform candidates from a different background, given similar linguistic competence.

There is still a problem, however, in defining the boundaries of specific context areas (Cumming, 2001; Davies, 2001; Elder, 2001). It appears to be the case that while we can identify particular aspects of language use as being specific to a given context (such as vocabulary, syntax, rhetorical organisation), we cannot readily identify exact limits to the language that is used in that context. This is because there are no 'exact limits'. Business language, like scientific or medical language is situated within and interacts with the *general language domain*, a domain that cannot, by its very nature, be rigidly defined.

4. Authenticity

Though Douglas (2000) built his definition of what makes a test 'specific' around the notions of situational and interactional authenticity, he later (Douglas, 2001) pointed to some difficulties in operationalising such a definition. The notion of situational authenticity is relatively easy to conceptualise. Situational authenticity refers to the accurate reflection in the test design of the conditions of linguistic performance from the language use domain. Tests such as that for air traffic controllers described by Teasdale (1994) are as close as we can get to a completely situationally authentic test.

The opposite to this would be the relative situational inauthenticity of the MATHSPEAK test referred to by Elder (2001) where there is no attempt made to replicate the teaching context to which it is designed to be generalised.

Though the common view of interactional authenticity which is that the test should result in an interaction between the task and the relevant language ability is clear enough, to my knowledge there has not been a significant contribution to its operationalisation. Test developers and researchers tend to rely on anecdotal evidence or 'expert' judgements to make decisions on the interactional authenticity of a test task.

So, critics of an LSP approach to language testing have raised genuine concerns regarding the distinguishability of distinct 'specific purpose' contexts, authenticity, and the impact on test performance of non-language factors – not just for LSP testing but for language testing in general.

5. Towards a theoretical conceptualisation of business language tests

The main thrust of this paper has been that it is not helpful to take the view that tests can only be seen as being 'specific purpose' if they are very narrowly focused on a particular 'purpose' area and are representative of, to borrow McNamara's (1996) expression, a 'strong' view of specific purpose testing. Instead there are a number of perspectives related to 'specific purpose' tests that offer a not incompatible expansion to the definition of SP tests offered by Douglas (2000:19). These include the following:

1. As all tests are in some way 'specific', it is best to think of all language tests as being placed somewhere on a continuum of specificity, from the broad general purpose test (such as the Certificate of Proficiency in English) to the highly specific test for air traffic controllers described by Teasdale (1994). This continuum can be visualised as shown in Figure 1.

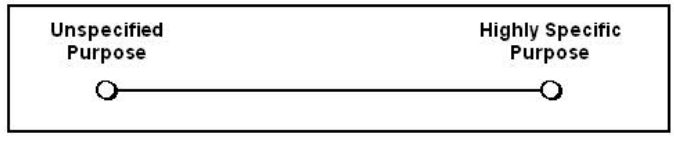


Figure 1. A view of test specificity.

2. Very highly specific tests tend to be very poor in terms of generalizability, while the opposite can be said of non-specific tests. There is not a binary choice in operation here, and if we accept that tests can be developed along a specificity continuum, then it logically follows that a test which appears to be placed somewhere other than the extremes of the continuum will have the potential to be either more or less generalizable.

3. Where a test is situated closer and closer to the more highly specific end of the continuum, the focus on *situational* authenticity also changes. That is, a highly specific test will most closely reflect the 'real world' situation or context, while a more general, less specific test will be less likely to do so (though it is not impossible that a specific context might be exploited in a test of general proficiency). In other words, a highly specific test will typically demonstrate *situational* authenticity.
4. Since we are essentially focused on tests of language, the aim of any specific purpose language test is to attempt to say something about a candidate's language ability within the specific context of interest. Therefore, the extent to which a test task engages a candidate's underlying processing and language resources to the same degree as called for within the specific context domain indicates the degree of *interactional* authenticity of that test task.
5. The degree to which non-language factors impact on a candidate's test performance will reflect the degree of specificity of that test. Therefore, in a highly specific language test it may not be possible to separate the language from the specific event. Where such a test is called for (i.e. a 'strong' form of specific purpose tests) this should be recognised in the definition of the construct and as such the only possible way to assess language performance should be within performance in the event, using, for example, the type of 'indigenous' assessment rubrics or scales suggested by Jacoby and McNamara (1999) and developed by Abdul Raof (2002).

It is clear from these five points that the notion of 'degree of specificity' is central to any definition of a specific purpose language test – since the impact of other factors will vary, depending on the positioning of a test along a specificity continuum.

6. Locating specificity

The notion of specificity, if it is to be of practical use to the test developer, must be tied to an understanding of test validity. One such perception of test validation is suggested by the series frameworks for all four skills presented by Weir (2005). In these frameworks, validity is seen from a socio-cognitive perspective (see Figure 2 for a summary of Weir's approach).

In this outline, we can see that there are a number of elements, each of which should be attended to by the test developer. Evidence is required at each level, in order to make validity claims for a test. I have added to the framework by highlighting the fact that the test taker can be described in terms of a number of characteristics (physical/physiological; psychological; and experiential) and by the internal processing (unique to the individual) which takes place during test performance. The test can be described in terms of its context validity and in terms of the potential for successful test

tasks to utilise appropriate processing. It is this notion of what Weir (2005) calls *theory-based validity* that forms the link between the test and the test taker.

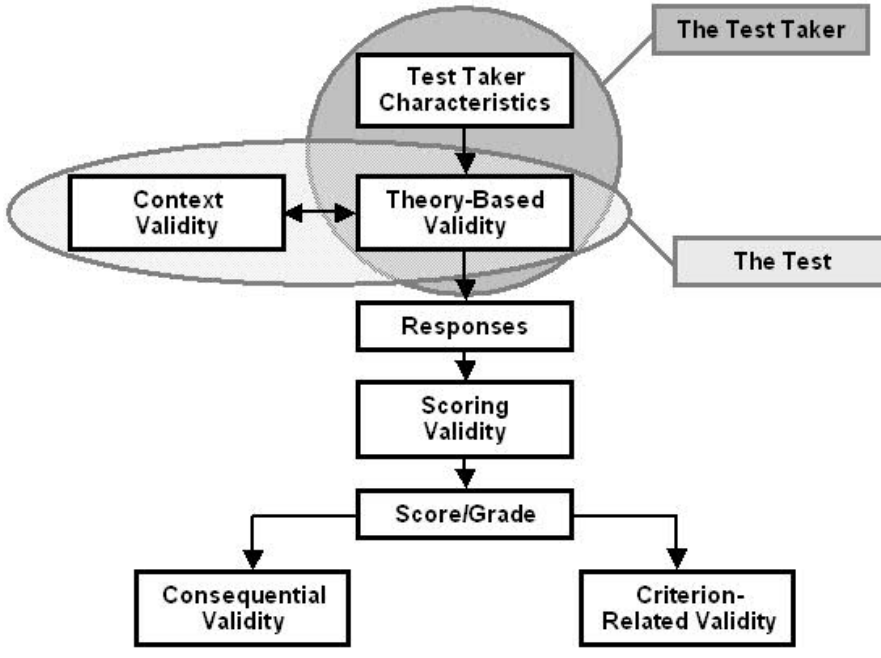


Figure 2. Format of Validation Frames (based on Weir, 2005).

O'Sullivan (2011) first attempted to describe an overarching socio-cognitive model (as opposed to Weir's frameworks, which were designed with specific language skills in mind). This underlying model has been further refined by O'Sullivan (2016), as well as by Chalhoub-Deville and O'Sullivan (in press). While the details within the model closely reflect those of Weir's (2005) frameworks, the approach has been reconceptualised to integrate consequence more fully into the whole process of test development and validation – in Weir's original it was seen as a post-test activity only.

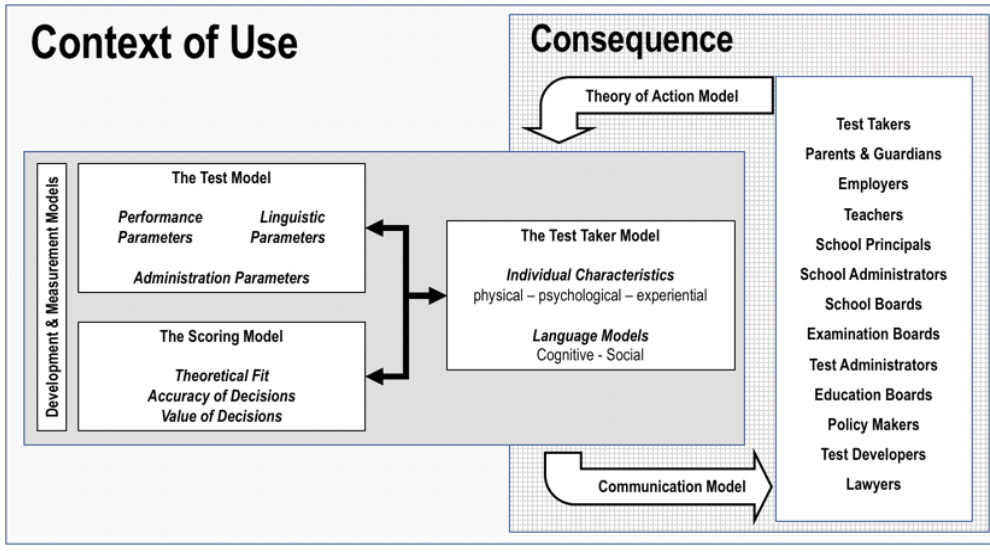


Figure 3. Socio-Cognitive Model as Integrated Arguments (from Chalhoub-Deville & O'Sullivan, in press).

The Test Model (see Figure 4) is concerned with aspects of the demands of the task and text, as well as detailing the test setting. In terms of the view of LSP tests offered here, it should become clear that when we are talking about test specificity, we are actually referring to test context, and this is expressed in the framework as being comprised of task and text demands.

When we consider the difficulty in defining language proficiency and use (for example the 'boundary' issue raised by Davies (2001) and Elder (2001), we can see that this aspect of validation is always going to be problematic. The operations and conditions suggested in the framework presented here are based on Weir (1993) and have been used with some success in test development projects for well over two decades, though they remain tentative in that there is no empirical evidence that these are the only operations and conditions applicable to a test of speaking (see Weir (2005) and O'Sullivan (2016) for a fuller description of the parameters included in the model).

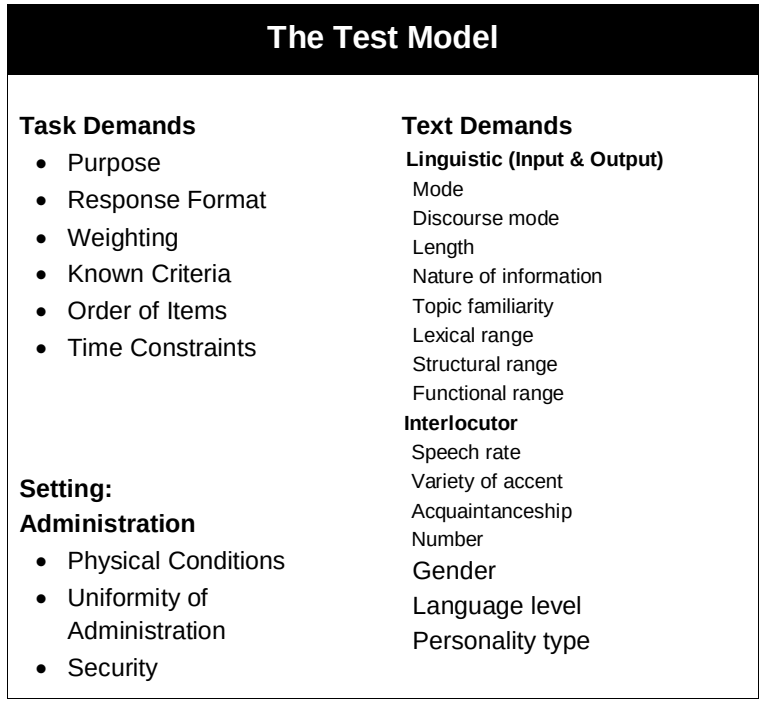


Figure 4. Aspects of Context Validity for Speaking, (Weir, 2005; O'Sullivan. 2016).

Test specificity might therefore be expressed as the degree to which the operationalisation of each of these demands can be considered to be uniquely related to a specific language use domain. In practice, this entails making value judgements of the degree of specificity along a continuum for each aspect of both task demands and text demands (see Figure 5). This may be seen as being too subjective a task to be of practical use. However, the real value of the exercise is in its breadth. Specificity is now seen as a multi-dimensional perspective of a test, and judgements are made on a systematic basis.

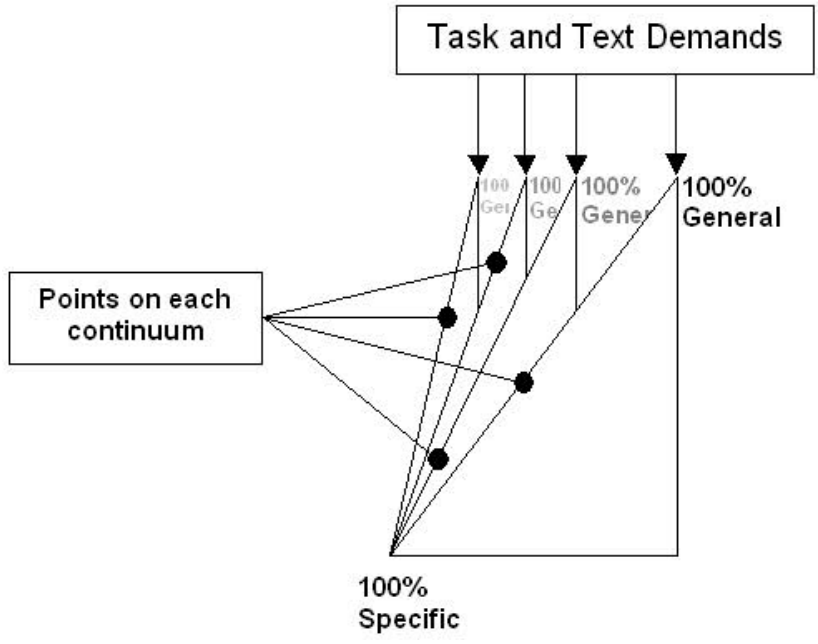


Figure 5. A Multi-Componential View of Specificity.

In order to demonstrate this multi-dimensional perspective, I undertook an experiment in which a group of language specialists were asked to take two test papers (of reading) and then make judgements on the papers based on a simple Likert scales-based instrument. The instructions to the specialists asked that they should try to decide where on the scales each of the two papers might lie, with 1 meaning very specific and 7 general - where an aspect was considered neutral it was decided that a rating of 4 should be awarded. The papers were taken from an LSP test, Business English Certificate Vantage (BEC), and a general proficiency test, the First Certificate in English, as these two tests are designed to allow for inferences to be made at the same Common European Framework of Reference (CEFR) level (B2). Figure 6 shows that there were clear differences seen by the specialists in terms of the task demands. This clearly different profile can be taken as empirical evidence of the distinguishability of LSP and general tests.

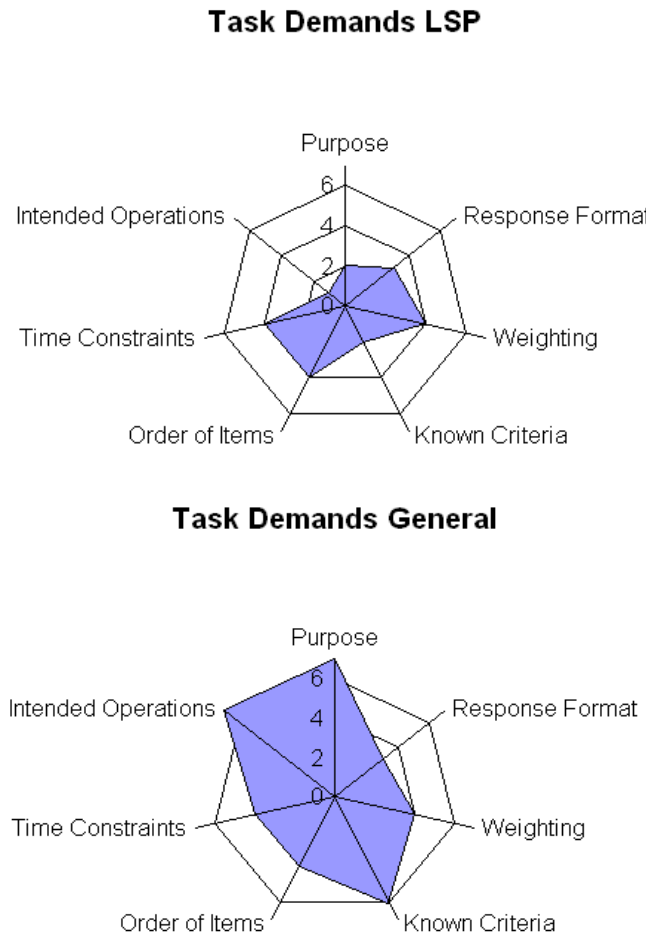


Figure 6. Differences in Task Demands between LSP and General Proficiency Test Papers.

When the participants were asked to repeat the exercise for the same papers, but this time with a focus on text demands, the differences are even more obvious (Figure 6).

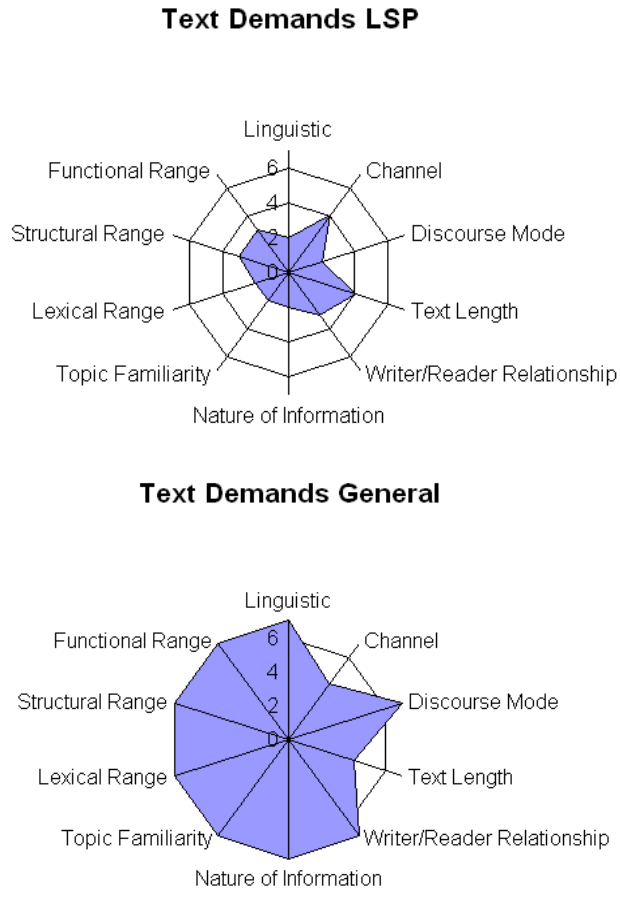


Figure 7. Differences in Text Demands between LSP and General Proficiency Test Papers.

The evidence from this, admittedly small study, suggests that judgements on the degree of specificity of an LSP test can be made in a systematic way. It also suggests that the notion of test specificity is closely linked to that of situational authenticity. Of course it could be argued that even a supposedly 'specific' test such as BEC, or even the 'highly specific' test described by Teasdale (1994) can never reach a position where the shaded area in the figures is minimised – indicating that the test has achieved a high degree of specificity from all perspectives. The evidence here supports the view that tests can never hope to do more than simulate authenticity, and intuition suggests that this same evidence will be found where other tests are analysed using the methodology suggested here – however highly specific the test developer claims it to be.

In the same way that the test model aspect of the socio-cognitive model can be used as the basis for establishing evidence of the specificity of a test, more evidence can be garnered from the other aspects of the model. While a lack of space prevents a thorough treatment of these, it should be clear that a similar approach to gathering data

as was used above could easily be applied to the other two aspects of the model, as will be briefly exemplified in the following two sections.

The Test Taker Model

Since we conceived of the construct as being located within the test taker (after all it is their language we are testing), we can look to the various aspects of the language model for similar evidence of specificity. Where we know that there are specific domain-specific usages of language these can be highlighted to demonstrate test specificity. The work of Hyland (2007, 2009, 2012, 2015) who demonstrates a range of domain specific language usage across academic writing, is particularly relevant here. It is also feasible that variables associated with the test taking population, in particular experiential characteristics, may also play a part in any specificity argument.

The Scoring Model

It is very clear that this model will be of particular importance when establishing test specificity. This is because we would expect that:

- Any rating criteria (scale or key) will be domain-specific
- Raters will be trained with the specific domain in mind
- Test decisions will have a specific domain as their basis
- Reporting will be done in a format that is meaningful to the specific domain
- Any criterion comparisons will be domain specific

7. Conclusion

I have tried to demonstrate in this contribution that tests of language for business purposes are different, both in their content and in their theoretical basis. I have offered a perspective on LSP testing that is supportable from both practical and theoretical perspectives and have added support to a definition of LSP tests presented in terms of authenticity (Douglas 2000). In doing this, I have come to the conclusion that the way we currently operationalise authenticity is somewhat naïve, and that authenticity is more complex than hitherto conceived.

All tests can be seen as lying on a specificity continuum, between the highly specific and the general purpose. This continuum is multi-componential and includes the twin aspects of authenticity – situational and interactional. A specific purpose test will be distinguishable from other tests (both specific and general purpose) in terms of the domain represented by the demands of its tasks and texts, and in terms of the cognitive processing it elicits.

When referring to tasks and content that ‘are authentically representative’ of the specific domain, Douglas (2000) is actually referring to the situational authenticity of the task and content. In light of the argument presented here I would suggest that his definition be revised to reflect an operationalisation of this form of authenticity, see Figure 8.

Situational Authenticity	
Douglas 2000	**Revised**
test tasks and content are authentically representative of tasks in the target situation	test tasks and content can be defined in terms of their position on a series of continua, each of which reflects an aspect of the demands that define the test task

Figure 8. A Socio-Cognitive Definition of Situational Authenticity.

Douglas (2000) refers to the interactional authenticity of a test task and content when he describes the three-way interaction between the task, the specific purpose content and the candidate's language ability. This aspect of his definition can be revised to reflect the broader understanding of the processing engaged in by candidates in a specific purpose test event, see Figure 9.

Interactional Authenticity	
Douglas 2000	**Revised**
an interaction between the test taker's language ability and specific purpose content knowledge, on one hand, and the test tasks on the other	an interaction of the test takers' executive resources and internal processes (i.e. their cognitive and meta-cognitive processing as well as their background and linguistic knowledge) and the context of the test task, as defined by the demands of that task

Figure 9. A Socio-Cognitive Definition of Interactional Authenticity.

The revised version of Douglas' definition can therefore be stated as:

> A specific purpose test is one in which test content and methods are derived from an analysis of a specific purpose target language use situation, so that test tasks and content can be defined in terms of their position on a series of continua, each of which reflects an aspect of the demands that define the test task, allowing for an interaction of the test takers' cognitive and meta-cognitive processing, their background and linguistic knowledge, the context of the test task, as defined by the demands of that task and finally the entire scoring model used within the test. Such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain.

Since any meaningful analysis of the language use situation will include a full description of the typical test taker, this definition allows the test developer to triangulate their definition of the construct to be tested. Consideration of the characteristics of test takers identified by O'Sullivan (2000a) alongside the cognitive and meta-cognitive demands of the test task will result in more meaningful and *authentic* tasks and scoring criteria, which in turn will allow us to make more supportable decisions based on test performance. Despite the drift away from the concept of construct validity in the literature observed by Chalhoub-Deville and O'Sullivan (in press), the reality of test development and validation is that construct definition lies at the heart of the entire process.

It is my hope that the more complete definition of authenticity proposed here, together with the expanded definition of specific purpose language tests will act to renew our understanding of the centrality of construct definition in test development and validation, and to revitalise our interest and scholarship in the area of specific purpose language testing.

References

Abdul Raof. A.H. (2002). *The production of a performance rating scale: An alternative methodology*. Unpublished PhD dissertation, The University of Reading, UK.

Alderson, J.C. & Urquhart, A.H. (1984). ESP Tests: The problem of student background discipline. In T. Culhane, C. Klein–Braley, D.K. Stevenson (Eds.). *Practice and Problems in Language Testing.* Essex: University of Essex. 1–13.

Alderson, J. C. & Urquhart, A.H. (1985). The effect of students' academic discipline on their performance in ESP reading tests. *Language Testing*, 2(2): 192–204.

Alderson, J.C. & Urquhart, A.H. (1988). This test is unfair: I'm not an economist. In P. Carrell, J. Devine & D. .Eskey (Eds.). *Interactive Approaches to Second Language Reading*. Cambridge: Cambridge University Press.168–182.

Alwright, J. & Alwright, R. (1977). An Approach to the Teaching of Medical English. In S. Holden (Ed.). *English for Specific Purposes*. Modern English Publications. 58–62.

Barber, C.L. (1962). Some Measurable Characteristics of Modern Scientific Prose. Reprinted in J. Swales, (Ed.). 1988. *Episodes in ESP*. New York: Prentice Hall, 1–16.

Beeching, K. (1997). French for specific purposes: The case for spoken corpora. *Applied Linguistics*, 18, 3: 374–394.

Berry, V. (1996). *Ethical considerations when assessing oral proficiency in pairs*. Paper presented at the Language Testing Research Colloquium, Tampere, Finland.

Berry, V. (1997). *Gender and personality as factors of interlocutor variability in oral performance tests*. Paper presented at the Language Testing Research Colloquium, Orlando Florida, USA.

Berry, V. (2004). *A study of the interaction between individual personality differences and oral performance test facets*. Unpublished PhD thesis, Kings College, The University of London.

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Us.* Cambridge Approaches to Linguistics. Cambridge: Cambridge University Press.

Brindley, G. (1984). The role of needs analysis in adult ESL programme design. In R.K. Johnson (Ed.). *The Second Language Curriculum.* Cambridge: Cambridge University Press. 63–79.

Brown, A. (1995). The Effect of Rater Variables in the Development of an Occupation Specific Language Performance Test. *Language Testing*, 12(1), 1–15.

Brown, A. (1998). *Interviewer style and candidate performance in the IELTS oral interview.* Paper presented at the Language Testing Research Colloquium. Monterey CA.

Brown, A. & Lumley, T. (1997). Interviewer Variability in Specific–Purpose Language Performance Tests. In A. Huhta, V. Kohonen, L. Kurki–Suonio, S. Luoma (Eds.) *Current Developments and Alternatives in Language Assessment*. Jyväskylä: University of Jyväskylä and University of Tampere. 137–150.

Canale, M. & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics*, 1, 1–47.

Carver, D. (1983). Some propositions about ESP. *ESP Journal*, 2: 131–137.

Chalhoub–Deville, M. & O'Sullivan, B. (in press). *Validity: Theoretical Development and Integrated Arguments*. Sheffield: Equinox.
Clapham, C.M. (1996). *The Development of IELTS: A Study of the Effect of Background Knowledge on Reading Comprehension*. Cambridge University Press, Cambridge.
Cumming, A. (2001). ESL/EFL instructors' practices for writing assessment: specific purposes or general purposes? *Language Testing*, 18, 2: 207–224.
Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing*, 18, 2: 133–147.
Dickerson, L. (1975). The learner's interlanguage as a system of variable rules. *TESOL Quarterly*, 9(4): 401–407.
Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.
Douglas, D. (2001). Language for specific purposes assessment criteria: where do they come from? *Language Testing*, 18, 2: 171–185.
Dudley–Evans, R. & St John, M.J. (1996). *Report on Business English: A review of research and published teaching materials*. TOEIC Research Report #2. Princeton NJ: Educational Testing Services.
Elder, C. (2001). Assessing the language proficiency of teachers: are there any border controls? *Language Testing*, 18, 2: 149–170.
Ellis, R. (1989). Sources of Intra–Learner Variability in Language. In S. Gass, C. Madden, D. Preston & L. Selinker. (Eds.). *Variation in Second Language Acquisition, Vol. 2: Psycholinguistic Issues*. Clevedon PA: Multilingual Matters. 22–45
Gledhill, C. (2000). The Discourse Function of Collocation in Research Article Introductions. *English for Specific Purposes*, 19: 115–135.
Hawkey, R. (1978). *English for Special Purposes*. London: British Council English Teaching Centre.
Hüllen, W. (1981a). Movements on Earth and in the Air: A study of certain verbs occurring in the language of international pilots. *ESP Journal*, 1, 2: 141–153.
Hüllen, W. (1981b). The teaching of English for special purposes: A linguistic view. In R. Freudenstein, et al (Eds.) *Language Incorporated: Teaching Foreign Languages in Industry*. Pergamon & Max Hueber Verlag. 57–71.
Hutchinson, T. & Walters, A. (1987). *English for Specific Purposes: A Learning–Centred Approach*. Cambridge: Cambridge University Press.
Hyland, K. (2007). *Writing in the Academy: Reputation, Education and Knowledge*. London: Institute of Education.
Hyland, K. (2009). *Academic Discourse: English in a Global Context*. London: Continuum.
Hyland, K. (2012). *Disciplinary Identities: Individuality and Community in Academic Discourse*. Cambridge: Cambridge University Press.
Hyland, K. (2015). *Academic Publishing: Issues and Challenges in the Construction of Knowledge*. Oxford: Cambridge University Press.
Hymes, D. H. (1972). On Communicative Competence. In J.B. Pride, & J. Holmes (Eds.) *Sociolinguistics: Selected Readings*. Harmondsworth, Middlesex: Penguin. 269–293.
Jacoby, S. & McNamara, T.F. (1999). Locating competence. *English for Specific Purposes*, 18(3), 213–241.
Johns, A. (1980). Cohesion in written business discourse: some contrasts. *ESP Journal*, 1(1), 35–44.
Kennedy, C. & Bolitho, R. (1984). *English for Specific Purposes*, Macmillan.
Lackstrom, J., Selinker, L., & Trimble, L. (1973). Technical rhetorical principles and grammatical choice. *TESOL Quarterly 7*. 127–136.
LCCIEB. (1972). *The Non–specialist Use of Foreign Languages in Industry and Commerce*. Sidcup: London Chamber of Commerce and Industry Examinations Board.
Munby, J. (1978). *Communicative Syllabus Design*. Cambridge: Cambridge University Press.
O'Sullivan, B. (2011). Language Testing. In J. Simpson (Ed.). *Routledge Handbook of Applied Linguistics*. Oxford: Routledge.

O'Sullivan, B. (2014). *Stakeholders and consequence in test development and validation*. Plenary Address at the Language Testing Forum, University of Southampton.
O'Sullivan, B. (2016). Validity: What is it and who is it for? In Yiu–nam Leung (Ed.). *Epoch Making in English Teaching and Learning: Evolution, Innovation, and Revolution*. Taipei: Crane Publishing Company Ltd.
O'Sullivan, B. & Weir, C. (2011). Language Testing and Validation. In B. O'Sullivan (Ed.). *Language Testing: Theory & Practice.*13–32. Oxford: Palgrave.
O'Sullivan, B. (1995). *Oral Language Testing: Does the Age of the Interlocutor make a Difference?* Unpublished MA Dissertation. University of Reading.
O'Sullivan, B. (2000a). *Towards a Model of Performance in Oral Language Testing*. Unpublished PhD Thesis, University of Reading, UK.
O'Sullivan, B. (2000b). Exploring Gender and Oral Proficiency Interview Performance. *System*, 28 (3) 2000: 373–386.
O'Sullivan, B. (2002). Learner Acquaintanceship and Oral Proficiency Test Pair–Task Performance. *Language Testing*, 19(3): 277–295.
O'Sullivan, B. (2006). *Issues in Testing Business English: The revision of the Business English certificates.* Studies in Language Testing Volume 17. Cambridge: Cambridge University Press
Porter, D. (1991a). Affective Factors in Language Testing. In J.C. Alderson, B. North (Eds.). *Language Testing in the 1990s*. London: Macmillan (Modern English Publications in association with The British Council), 32–40.
Porter, D. (1991b). Affective Factors in the Assessment of Oral Interaction: Gender and Status. In Sarinee Arnivan (Ed.). *Current Developments in Language Testing*. Singapore: SEAMEO Regional Language Centre. Anthology Series 25: 92–102
Porter, D. & Shen Shu Hung (1991). Gender, Status and Style in the Interview. *The Dolphin 21*, Aarhus University Press: 117–128.
Robinson, P.C. (1980). *English for Specific Purposes.* Oxford: Pergamon
Robinson, P.C. (1985). *Needs Analysis: From Product to Process.* Paper presented the 5th Annual ESP Symposium, Leuven, Belgium, August 1985.
Schmidt, M. (1980). Coordinate structures and language universals in interlanguage. *Language Learning*, 30: 397.
Schröder, K. (1981). Methods of Exploring Language Needs in Industry. In R. Freudenstein, et al. (Eds.). *Language Incorporated: Teaching Foreign Languages in Industry.* Pergamon & Max Hueber Verlag. pp. 43–54.
Selinker, L. & Douglas, D. (1985). Wrestling with 'Context' in Interlanguage Theory. *Applied Linguistics*, 6, 2: 190–204.
Smith, J. (1989). Topic and Variation in ITA Oral Proficiency. *English for Specific Purposes*, 8, 155–168.
Spolsky, B. (1995). *Measured Words*. Oxford: Oxford University Press.
Steffensen, M.S. & Joag Dev, C. (1984). Cultural Knowledge and Reading. In J.C. Alderson, & A.H. Urquhart (Eds.). 1984. *Reading in a Foreign Language*. London: Longman.
Swales, J. (1971). *Writing Scientific English*. London: Thomas Nelson.
Swales, J. (1984). ESP comes of age? – 21 years after 'Some Measurable Characteristics of Modern Scientific Prose'. *UNESCO Alsed – LSP Newsletter*, 7, 2 (19): 9–20.
Tarone, E. (1985). Variability in Interlanguage use: A study of style shifting in morphology and syntax. *Language Learning*, 35(3): 373–404
Tarone, E. (1988). *Variation in Interlanguage*. London: Edward Arnold.
Teasdale, A. (1994). Authenticity, validity, and task design for tests of well defined LSP domains. In R. Khoo, (Ed.). *LSP Problems & Prospects*. Singapore: SEAMEO RELC: 230–242.
Thurstun, J. & Candlin, C. (1998). Concordancing and the teaching of the vocabulary of Academic English. *English for Specific Purposes*, 17, 3: 267–280.
Weir, C.J. (1993). *Understanding and Developing Language Tests*. London: Longman

Weir, C.J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Oxford: Palgrave Macmillan.

Weir, C.J. & Milanovic, M. (Eds.). (2003). *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*. Cambridge: University of Cambridge Local Examination Syndicate & Cambridge University Press.

West, R. (1994). Needs analysis in language teaching. *Language Teaching*, 27(1), 1–16.

An appreciation of Sauli Takala's contribution to Council of Europe language projects

Johanna Panthier and Joe Sheils
Formerly of the Language Policy Division, Council of Europe, Strasbourg

Sauli's initial involvement with the Council of Europe can be traced back to the 1970s when he was the Finnish representative in the EUDISED project to develop an agreed European system for documenting and disseminating information on educational research and development work.

His close identification with the values and goals underpinning the Council's work in education is evident in his remarkable contribution to successive modern languages projects over four decades. He will be remembered in particular as a highly respected influential member of working groups that contributed to the development of the *Common European Framework of Reference for Languages: learning, teaching, assessment* (CEFR), and related tools developed to support its effective implementation.

Soon after publication of the CEFR he was instrumental in the organisation, by the Finnish authorities, of a forum in Helsinki which resulted in the Council of Europe launching a project concerned with appropriate linking of language examinations to the CEFR. Sauli, after coordinating this seminal event, inevitably became a key member of the small expert group that went on to develop the *Manual for Relating language examinations to the CEFR.*

Sauli's remarkable expertise was further demonstrated as editor of the *Reference Supplement to the Manual for Relating Language examinations to the Common European Framework of Reference for Languages,* where he was ably assisted not only by his fellow authors but also the late Dr Felianka Kaftandjieva (University of Sofia). The technical guidance offered by the authors ensures that users of the above *Manual* have the option of considering additional processes when relating their certificates and diplomas to the CEFR, taking into account quantitative and qualitative considerations and different approaches in standard setting. One of Sauli's key concerns as editor was to ensure that the authors' contributions would be as accessible as possible, keeping technical language to a minimum and providing concrete examples, figures and tables to illustrate the processes involved. Always conscious that highly demanding subject matter cannot be simplified beyond a certain point without risking oversimplification, he was careful to sound a note of caution concerning oversimplifications that many "rules of thumb" imply.

No matter how busy, Sauli always made himself available to assist the Council of Europe, not only with the wide range of activities related to the CEFR, but also with curriculum quality and renewal more generally. We were particularly grateful for his expert assistance and leadership in a project to support the elaboration and implementation of new curricula for modern languages in Bosnia and Herzegovina. Sauli's quest for quality meant that he was a natural choice as consultant for the ECML colloquium on "Quality in Language Testing". He was always keen to support initiatives that aimed at "Empowering Language Professionals".

We remember how Sauli's interventions at meetings in Strasbourg were always so thoughtfully and politely expressed, and his unassuming manner no doubt contributed significantly to the mutual respect and collegiality that characterised our expert meetings.

The last completed Council of Europe activity in which Sauli was centrally involved, as project co-coordinator with Neus Figueras, aimed to gather illustrative tasks that could help users relate locally relevant test items to the CEFR levels, while gaining insights into the development of items that can claim to be related to CEFR levels. More than a decade had elapsed since the publication of the CD which included samples of listening and reading items to exemplify the procedures outlined in the Manual for Relating Language Examinations to the CEFR. This new project was undertaken in response to numerous requests for additional CEFR exemplar test tasks and items to facilitate competent assessment of proficiency in reading and listening comprehension.

The new collection of exemplar test tasks and items was made available on the web in 2017, thanks to a team of experts working under the guidance of Neus and Sauli (both of whom are among the co-authors of the Manual and responsible for the CD). These additional exemplars complement the earlier CD and the samples of oral and written production previously made available by the Council of Europe (in five languages) for relating language examinations to the CEFR levels.

Sadly, in the course of this project Sauli experienced serious health problems, culminating in a major heart operation. In spite of this setback and in keeping with his unstinting commitment and dedication, he remained in contact with his co-coordinator and with the Secretariat at the Language Policy Division until the successful completion of the project.

It is hardly surprising, therefore, that Sauli remained involved in Council of Europe work to the very end. He contributed to early work on the development of the *CEFR-Companion Volume* in his role as a member of the six-person sounding board set up to support the authoring group. He brought the same calm, insightful, critically constructive approach to bear on this challenging new venture that characterised his contribution to all projects. The last time we saw Sauli was when he attended a CEFR seminar at the Council of Europe in Strasbourg, in June 2016. We were shocked and saddened to learn of his tragic death in February 2017.

In view of Sauli's universally recognised expertise and experience concerning the CEFR, it can be safely assumed that the CEFR-CV project could not fail to have

been enriched by his continued collaboration, no doubt guided, as always, by his inspiring commitment to the shared values promoted by the Council of Europe.

We wish to record our deep appreciation and gratitude for the remarkable contribution made by Sauli Takala to the promotion of the values and goals underpinning the Council of Europe's work in the field of language education over four decades.

A note on changing attitudes to linguistic errors in learner language in English teaching in Norway

Aud Marit Simensen
University of Oslo

1. Introduction

Sauli Takala, a dedicated applied linguist and a dear friend, was a recognised expert in language testing and assessment. For that reason. I have chosen a topic where the question of assessment is central.

Practicing teachers of English as a foreign language have probably always wanted their students to obtain the highest possible degree of linguistically correct English, whatever norm they have been aiming at. However, views have shifted over time in the educational community as to how to handle linguistic errors in practical teaching as well as in assessment. Furthermore, opinions have differed from one time to another in the history of English teaching with regard to how to account for the linguistic errors[39] that actually occur.

Over time several academic disciplines have taken an interest in and responsibility for teaching and learning foreign languages. This also applies to questions of linguistic errors in learner language. In discussions of foreign language teaching and learning these disciplines are sometimes referred to as "parent disciplines" (for a more detailed discussion of this concept and its practical consequences, see Simensen 2007).

In the following, I will give a glimpse only of how teachers of English as a foreign language (EFL) in compulsory schooling in Norway have been instructed or recommended in national mandatory documents to recognise and deal with such questions. A central and interesting phase in this respect was one period of the 20th century, when teaching and assessment were guided by an experimental curriculum, *Læreplan for forsøk med 9-årig skole. Forsøk og reform i skolen*, nr. 5 of 1960 (L-60). This was analysed in one part of my PhD study (Simensen, 1988 and Simensen, forthcoming 2019). To my knowledge no comprehensive research of assessment practice in EFL at compulsory school level exists for the period after the 1980s.

[39] The distinction often made between mistakes and errors is not made here.

This phase was noteworthy also for other reasons: it introduced centralised final school leaving exams in the written skills for this level of compulsory schooling (for details see Simensen 1988) and it was exceptionally well documented in central official documents for English teaching and assessment, such as key subject curricula and corresponding assessment documents, including the yearly written test batteries. To some extent these documents also stand out since they in a remarkably clear way reflect correspondence between shifts of central conceptions about learners' linguistic errors in academic disciplines and in educational fields.

It should be kept in mind that the introduction of EFL in compulsory schools in Norway was a gradual process. It started around the end of the 19th century with English teaching in a few schools only, normally as a voluntary subject for the pupils and essentially thanks to local initiatives in districts along the southern coast of the country (see Gundem 1989). The increase from seven to nine years of compulsory education was also a gradual process starting on an experimental basis and continuing as such up to the passing in 1969 of *Lov om grunnskolen av 13. Juni 1969* which introduced the comprehensive school system in Norway and English as a compulsory school subject for all pupils *nationwide*. And from a general educational as well as a subject specific point of view it was implemented in stages and finalised five years later in the curriculum *Mønsterplan for grunnskolen* (1974; M-74).

What follows will start with approximate suggestions of two periods of English teaching in Norway followed by a selection of key words about especially influential activities, ideas, theories etc. in the parent disciplines, under the headlines "factors of influence". Then I will give an account of some significant instructions for EFL included in the aforementioned national mandatory documents for each period. These sections will carry the headlines "EFL in Norway".

2. From the end of the 19th century up to the middle of the 20th: a focus on speech, correct pronunciation and direct associations

2.1 Factors of influence

The most important amendment of the well-known Reform Movement in Europe was the shift of focus from written language to focus on speech. The impact of the movement was largely due to the fact that pioneer linguists at the time had provided necessary practical tools for the teaching of oral skills: a phonetic alphabet and phonetically transcribed texts (see Howatt with Widdowson's (2004), authoritative book on English language teaching). In addition, the academic discipline of psychology provided the idea of establishing direct associations in learning implying that words in the foreign language in practical teaching should be associated *directly* with the relevant thing, idea, act etc. talked about and not with the more or less equivalent words in the learner's first language. This led to the development of *the direct method teaching theory*.

2.2 EFL in Norway

The shift of focus outlined above was a slow process in Norway like in many other countries in Europe. Poor oral skills in English among teachers was the rule, not the exception. In Norway this meant, for example, that for some time textbooks based on the traditional and well-established grammar-translation teaching theory had to be used alongside textbooks based on direct method teaching.

The first subject curriculum with an executive function for English in compulsory school, was *Normalplan for byfolkeskolen* (N-39). This curriculum referred to the direct method teaching theory and had "good pronunciation" as one of four major objectives. An introductory course in phonetics and a study of phonetically transcribed texts were prescribed. Prescriptions were also given for the work of the teachers: "The pronunciation must throughout the whole course be given the most meticulous attention." Furthermore, teachers were advised "to be extremely conscientious to correct even the smallest errors" and "a correct pronunciation" was strongly emphasised in teaching. It added that incessant practice will convert what is taught into unfailing habit. (N-39, pp.236- 238; my translations).

3. From the middle of the 20th century towards the end of it, including the experimental period: contrastive analysis, language habits, language acquisition device, natural order and interlanguage

3.1 Factors of influence

The interest in language teaching and learning among linguists and applied linguists at this time was increasing. This had great consequences for the changes to come. Among other things, it meant describing languages in terms of sentence structures (structuralism). For educational purposes, professionals in these disciplines were mobilised to apply the new approach to different languages as well as to compare them (contrastive analysis). The purpose was first and foremost to find out on which points the foreign language differed from the learner's first language. "The contrastive analysis hypothesis" at the time predicted that differences between the two languages involved could result in linguistic errors in the learner's language. However, this could be counteracted in teaching by means of specially constructed texts and multiple exercises based on points of differences between the languages. The faith in the effect of contrastive analysis was once expressed in the following way: "Like sin, error is to be avoided and its influence overcome, but its presence is to be expected" (Brooks 1960, p. 58).

The idea of associations in learning from the previous teaching theory was replaced by the theory of establishing language habits, as expressed in the following way in one of the most celebrated slogans at the time "Language is a set of habits" (Moulton 1961, pp. 86-90); cp. Behaviourism). A great number of exercises were

considered necessary and strict control in teaching was indispensable to ensure that learners only produced correct responses. Ideas from structuralism and concepts such as habit formation led to the development of *the audio-lingual teaching theory*.

The most radical change of theory to come after audiolingualism was Noam Chomsky's notion that human beings are born with an innate language learning ability, "a language acquisition device", which claimed that language develops through *exposure* to meaningful language (for example Chomsky 1959). Stephen Krashen (1982) took this idea a step further in the natural order hypothesis of his Monitor theory. Among other things, this hypothesis maintained that the acquisition of grammatical structures proceeds in a predictable order and that certain errors in learner language might actually signify that the learner had advanced one step further towards a fully developed and correct target language. The steps the learners went through were sometimes called *interlanguages*. This completely new mind-set challenged the leading ideas of just a couple of decades before. And it is unlikely that articles with titles like the following passed by unnoticed in foreign language teaching communities: "You can't learn without goofing" (Dulay & Burt, 1974, p. 95) and "Should we count errors or measure success?" (Enkvist, 1973, p.16).

3.2 EFL in Norway

As noted above, *Læreplan for forsøk med 9-årig skole. Forsøk og reform i skolen*, nr. 5 of 1960 (L-60) was an experimental curriculum formally in effect up to the new law in 1969 (*Lov av 13. juni om grunnskolen*). However, experimental or not, aspects of practical teaching of English described in L-60 remained more or less in accordance with the direct method teaching theory of the first part of the 20th century.

Part of the experiment was to develop new assessment systems as explained in the corresponding guideline for assessment, *Evaluering i 9-årig skole. Metodisk veiledning* of 1964, as well as in later documents with a similar function (see *Evaluering i 9-årig skole. Avgangsprøva* 1964-1970, *Evaluering i grunnskolen. Avgangsprøva 1971-1973*, and *Evaluering i grunnskolen. Avgangsprøva* 1974-1986).

In general, these documents included prescriptions for assessment of the pupils' written language only (for details about final exams and types of tests given, see Simensen, 1988). They were clearly influenced by new ideas as discussed in the previous section. Together with a succession of new subject curriculums to come during the last two decades of the 20th century, they reflected changing attitudes to errors in learner language in the Norwegian educational community.

The guideline documents for assessment to follow the first guideline of 1964, *Metodisk veiledning*, became leading "trendsetter" texts as to assessment of learner language with linguistic errors. At the start of this development a general principle was that positive as well as negative aspects of the pupils' language at exams should be noticed. Assessors were furthermore advised to distinguish between essential and *non-essential errors*. A few years later non-essential errors were specified as formal errors that do not distort (*fordreier*) the meaning in the pupils' written language (*Norsk Skole*,

nr. 8, 1967, p. 277). The advice was that such errors should not be given too much emphasis in assessment. In documents to come assessors were even given examples of sentences with "non-essential" errors as well as "essential" errors, accompanied by scales with points to be given for each type (see for example *Evaluering … Avgangsprøva* 1974-1986, p. 30).

An interesting feature of this development was the instruction from 1965 that recommended giving the pupils credit if the *writing* reflected a good *oral* command of the language. In 1970 this was called "aural assessment" (*auditiv vurdering*) (*Evaluering … Avgangsprøva* 1964-1970, p. 49). This was regularly referred to in later documents. It is tempting to interpret this as an effort to compensate for less or no focus on assessing oral skills at final school-leaving exams (for a different type of compensation, see Simensen 1988, for example p. 50).

The subject curriculum published in 1974 (M-74) and introduced above, maintained, for example, that speech habits are most efficiently established through the production of correct responses. The teacher was therefore advised to direct controlled oral exercises in such a way that errors could be avoided. These and other statements reflected central tenets of the audiolingual teaching method. However, at the same time, it was emphasised that a fear of making errors must be avoided and that exercises involving less control should also be used for the expression of meaning. Actually, M-74 underlined the conception of language as a means to contact with other people, and the concept of *communication* was referred to for the first time in an English subject curriculum in Norway. Several examples of the same nature indicated that the theoretical bases of the audiolingual method had already been questioned before the most audiolingually-oriented subject curriculum in Norway, M-74, was published.

An updated version of the M-74, *Mønsterplan for grunnskolen* (M-87), was implemented in 1987. In this document, a strict control to avoid errors in the learners' language was no longer regarded as necessary. On the contrary, teachers were advised to help their pupils to develop a "constructive attitude to language errors when using English". Moreover, pupils should be taught that "instead of being afraid to make mistakes, they must understand that they can learn from their mistakes" (quote from the English version of M-87, *Curriculum Guidelines for Compulsory Education in Norway* (1990), p.223). This attitude is even taken one step further in the next subject curriculum, *Læreplanverket for den 10-årige grunnskolen* of 1997 (L-97), the last subject curriculum of the 20th century and the last to be mentioned here but not further commented. The following formulation in 1997 was ground-breaking: "Errors can be seen as signs of learning" (*Curriculum for the 10-year Compulsory School in Norway* (1999), p. 242, the English version of L-97). Hopefully, new research will show what happened in assessment practice at the end of the 20th century.

4. Conclusion

In this note, I have summarised aspects of the development of changing attitudes to linguistic errors in the language of learners of English in the educational community in Norway. Especially from 1960 onwards a number of fundamental changes took place. The study which constituted the basis of the present note had for example shown that there was more correspondence than I had expected between ideas, theories and tools in parent disciplines and in EFL in Norway, even to the extent of similar verbal expressions used in academic sources and in educational sources.

References

Brooks, N. (1960). *Language and Language Learning. Theory and Practice.* New York: Hartcourt, Brace and World.

Chomsky, N. (1959). A review of B.F. Skinner's Verbal Behaviour. *Language*, XXXV, 1, 26-58.

Curriculum for the 10-year Compulsory School in Norway. (1999). Oslo: The Royal Ministry of Education, Research and Church Affairs.

Curriculum Guidelines for Compulsory Education in Norway. (1990). Oslo: Aschehoug.

Dulay, H.C. & M.K. Burt. (1974). You can't learn without goofing: an analysis of children's second language learning strategies. In J.C. Richards (Ed.), Error Analysis: Perspectives on Second Language Acquisition. (pp. 95-123). London: Longman.

Enkvist, N.E. (1973). Should we count errors or measure success? In J. Svartvik (Ed.). (1973). Errata. Papers in error analysis. (pp. 16-23). Lund: CWK Gleerup.

Evaluering i grunnskolen. Avgangsprøva 1971 -1973. Oslo: Grunnskolerådet.

Evaluering i 9-årig skole. Avgangsprøva 1964 -1970. Oslo: Universitetsforlaget.

Evaluering i 9-årig skole. Avgangsprøva 1974 -1986. Oslo: Universitetsforlaget.

Evaluering i 9-årig skole. Metodisk veiledning. (1964). Oslo: Folkeskolerådet.

Gundem, B.B. (1989). *Engelskfaget i folkeskolen. Påvirkning og gjennomslag fra 1870-årene til først på 1970-tallet.* Oslo: Universitetsforlaget.

Howatt, A.P.R. with H.G. Widdowson. (2004). *A History of English Language Teaching*. (2nd ed.). Oxford: Oxford University Press.

Krashen, S.D. (1982). *Principles and Practice in Second Language Acquisition*. New York: Prentice Hall.

Lov av 13.juni 1969 om grunnskolen. (1969).Oslo: Grøndahl & Søn.

Læreplan for forsøk med 9-årig skole. Forsøk og reform i skolen, nr. 5. (1960). Oslo: Forsøksrådet for skoleverket (distributed by Aschehoug).

Læreplanverket for den 10-årige grunnskolen. (1997). Oslo: Kirke-, utdannings- og forskningsdepartementet.

Moulton, W. G. (1961). Linguistics and Language Teaching in the United States 1940–1960. In C. Mohrman, A. Sommerfeldt, & J. Whatmough (Eds.), *Trends in European and American Linguistics 1930–1960* (pp. 83–109). Antwerp: Spectrum Publishers.

Mønsterplan for grunnskolen. (1974). Oslo: Aschehoug Forlag.

Mønsterplan for grunnskolen. (1987). Oslo: Aschehoug Forlag.

Normalplan for byfolkeskolen. (1939). Oslo: Aschehoug & Co.

Norsk Skole, nr. 8, 1967, p. 277.

Simensen, A.M. (1988). *Et kvart århundre med evaluering i skolefaget engelsk* (Part of PhD thesis, 1988). Trondheim: University of Trondheim.

Simensen, A.M. (2007). *Teaching a Foreign Language. Principles and Procedures* (2nd ed.). Bergen: Fagbokforlaget.

Simensen, A.M. (forthcoming 2019). PhD revisited: English in compulsory school. Aims and content. In U.E. Rindal & L.M. Brevik (eds.), *English didactics in Norway – 30 years of doctoral research.* Oslo: Universitetsforlaget. ISBN 9788215030746.

Standard setting for writing and speaking: The Saint Petersburg experience

Norman Verhelst
Eurometrics, The Netherlands

Neus Figueras
University of Barcelona/Departament d'Ensenyament, Catalonia, Spain

Elena Prokhorova
Formerly of the Saint Petersburg State University, Russia

Sauli Takala
Formerly of the University of Jyväskylä, Finland

Tatiana Timofeeva
Formerly of the Saint Petersburg State University, Russia

Preamble by Norman Verhelst

It must have been at the EALTA conference of 2012 in Innsbrück that EALTA's wandering ambassador Dianne Wall contacted Sauli Takala, Neus Figueras and myself (as members of the authoring group of the Manual to link examinations to the CEFR) with the request to help the Language Testing Centre of Saint Petersburg State University with a linking project. Although we agreed immediately in principle, the seeking of contact in the beginning was hesitant and slow, and it was only at the next EALTA conference in Istanbul in 2013 that we met Elena Prokhorova and Tatiana Timofeeva, responsible for the Language Testing Centre, and that we could start to make plans. This resulted in two important meetings in 2014: one in June for setting standards for Listening, Reading and Use of English and one in November for setting standards for Writing and Speaking.

The present chapter is about the latter standard setting with an almost exclusive focus on the techniques used. Although it is theoretically possible to present this work as a one-author article, this would distort reality to a large degree. The preparatory work and the success of the project depended on thorough discussions between people with a very different background in testing, a simple but sound introduction in statistical and psychometric techniques and a good understanding of the CEFR. Also the three of us, as consultants had to work closely together to guarantee a consistent approach to the work to be done. It was due to a close collaboration and mutual understanding of all people involved in the standard setting process that the project was successful. Therefore the list of co-authors is long, and could even have been longer and completed by all the members of the panel, internal and external, who did their judgmental job in a professional way.

This project was the last one where I collaborated with my dear colleague and friend Sauli Takala, with whom I have had numerous discussions on language testing as well as on statistics and psychometrics. We both were convinced that interdisciplinary collaboration is the key to real progress in science.

1. Introduction

The Second Certificate Test in English (further - the University Test) was developed by the Saint Petersburg State University (further - the SPSU) Language Testing Centre as an exit test in English for Bachelor level students of non-linguistic faculties. The University Test was piloted in 2008, 2009, 2010 and finally in 2011 was implemented as the required exit test.

In 2012, the SPSU authorities made a decision to carry out the University Test evaluation using international experts to confirm its validity and reliability. The new SPSU external assessment system was analysed using a framework designed by the Association of Language Testers in Europe (ALTE) – *17 Minimum Standards*. The examination analysis resulted in the conclusion that the University Test is a test in General Academic English that is carefully thought out, reliable and well correlated with its educational aims. The Language Testing Centre not only ensured that the test content and structure complied with the aims and context in which it is used, and took into account the users' needs, but it also ensured that a balance was found between the validity, reliability and practicality of the external assessment system. The test requirements set forth in the University Test Specification are both realistic and clearly defined. The test itself follows the Specification and checks language competences. All the examination cycle stages are developed in detail, are practical and form an integrated and logical assessment system.

These conclusions allowed the SPSU authorities to decide on the second stage of the University Test evaluation – an analysis with international experts, which aimed to link it to the B2 CEFR level.

The main objective of linking the University Test to the CEFR, which started in November 2013, was to provide valid and reliable confirmation of the fact that the proficiency in English of those who have passed the University Test is the B2 CEFR level.

For the productive skills, Writing and Speaking, it was agreed between the Test Centre and the consultants to use the Body of Work method (Kingston, Kahl, Sweeny and Bay, 2001; Kingston and Tiemann, 2012; Cizek and Bunch, 2007, Chapter 9 and Council of Europe, 2009, Section 6.6). In preparation of the standard setting event the Language Testing Centre took care in the necessary preparatory activities of Specification and Standardisation and great care was given to the study and understanding of the CEFR.

In this chapter attention will be given to the technical aspects of the Body of Work method as it was implemented in the standard setting event of November 2014. The point of view taken is to report in quite great detail on the use of this method for

productive skills on the one hand, but on the other hand to serve as a guide for language testers who want to use the method in their own work. A number of features not presented in the cited literature will be discussed in detail.

In the next section the data collection design is discussed in relation to the practical constraints and to the validity of the standards. Then follows a section describing the statistical model used, the procedure of collecting the data and the main results. In a separate section some novel features of the model are discussed. They have to do with differences between panel members with respect to leniency and discrimination.

2. The data collection design

A central question in all standard setting procedures is the validity question: how can one build an argument that convinces the expert and general public that the standard(s) as set is/are really justified as the best delimiting border between passing and failing, or as is the case here, between being at the CEFR level B2 (or higher) and not having reached the B2 level. In this respect two decisions were taken from the very beginning:

1. The available panel for the standard setting consisted of people working at the Language Testing Centre and consequently were involved in the construction of the tests. As a safeguard against in-crowd culture, it was decided to try to add an external panel to the internal one, and to give special attention to systematic differences in judgments between the two panels.
2. The Language Testing Centre has since its beginning always worked along well established rules: using the specifications for the productive skills, tasks were constructed for every examination session (two per academic year) along the same lines and the scoring rubrics were constant over time, such that it was (allegedly) justified to consider test scores across examinations as comparable (e.g., a score of 23 on examination 1 reveals a higher skill than a score of 22 on examination 2). If this is really the case, then standards set for several examinations (creating very diverse 'bodies-of-work') should lead to the same standard or cut score.

In contrast to these two – reasonably sounding – requirements, are the practical constraints of the Body of Work method. In this method a panel member reads the script of the student to be judged (Writing) or hears a recording of his/her oral performance. In both cases the task for the panel member is to answer the following question: **'is a person who [writes | speaks] like this at level B2 of the CEFR (or above)? Yes or No?'** As the question is simple enough, the task for the panel members is not that simple: for the student an assignment in Writing or Speaking consisted of two tasks, and a conscientious fulfillment of the task for the panel members required to read or listen to the performance of the student in both tasks and to give one global answer to the question of the level, sometimes requiring a thoughtful weighting of weak and strong points in either of the two tasks. This practical consideration lead to an

important decision: as only a single day (per skill) for the standard setting event with the internal panel could be planned, the number of student works to be judged was set to 24 for Writing and to 16 for speaking. At the time of this planning, an internal panel of 13 judges was foreseen. A second practical decision was that each student's work was to be judged by 4 different judges.

As to the validity arguments, it was decided to use two different assignments, i.e., two pairs of two tasks, to find out how stable the standards were across assignments. Henceforth the two assignments will be labeled as set A and set B.

To have judgments beyond the inner circle of collaborators of the Language Test Centre, a number of language testing experts would be asked to do the judgments, without attending physically the standard setting event in Saint Petersburg.

The Body of Work method of standard setting essentially consists of two rounds, a *range finding* and a *pinpointing* round. The purpose of the former round is to identify a preliminary location of the standard, and in the second round judgments are collected from students' works in a more or less wide neighborhood around this preliminary standard. The preparation of the folders with works for the second round can only be done after the first round is finished and the answers are analysed. In a practical sense this means that there is a serious time gap between the first and second round (especially if an audio folder must be prepared for the Speaking test) or that a substantial number of second round folders must be prepared to cope with the many different possible outcomes of the first round. Both possibilities were deemed impractical and too expensive, and therefore it was decided to skip the range finding round, and to trust the experience of the Language Test Centre who had previously set standards, i.e. cut scores in several examinations, be it without a formal standard setting procedure.

To fulfil the requirements for a reasonable workload, the number of performances should be[40]

$$\frac{13\times 16}{4} = 52 \text{ for Speaking and}$$
$$\frac{13\times 24}{4} = 78 \text{ for Writing.}$$

An essential feature of the Body of Work method is that performances are selected in the designated range which are as uniformly distributed as possible. For Writing students did two tasks, each worth 20 points and one speaking task, worth 25 points. The scoring rubrics for speaking are given in appendix 1; those for Writing in appendix 2. Although Cizek and Bunch (2007) stress the importance of a range finding round to find the approximate cut-score, such a round was skipped for two reasons: first, it would have involved more time than was available for the standard setting procedure and, second, the statistical analysis planned was different from the one used by Cizek

[40] To understand the formula, here is the reasoning for Speaking: there are 13 judges and each judge gives 16 ratings. So in total there will be 13 x 16 judgments, but not as many students, since each student is judged by 4 judges. Therefore the total number of students involved is as given in the text.

and Bunch, as will be discussed further down. In our approach it was decided to use a score range that covered a high percentage of the obtained scores and chosen in such a way that the standard would be roughly half-way that range, where we trusted the experience of the colleagues from the Language Centre. For Writing the scores ranged from 16 to 35 and for Speaking from 11 to 20, giving 20 and 10 different score values, respectively. Of course, 10 and 20 do not divide 52 and 78, so an equal frequency of each score point is not possible, but an effort should be made to approximate the uniform distribution as well as possible. As an example, Table 1 displays the proposed distribution of scores for Speaking.

Table 1. Proposed frequency distribution of scores for Speaking.

score	**set A**	**set B**
11	2	2
12	3	3
13	2	2
14	3	3
15	3	3
16	3	3
17	2	2
18	3	3
19	2	2
20	3	3
total	**26**	**26**

In practice, the selection from the available recordings should follow the proposed frequencies as shown in Table 1, but otherwise be random. This means (see Table 1) that from all the available recordings having a score of 16, three should be chosen at random. There is one exception allowed: if the performance is very heterogeneous across the tasks, then it is not eligible for the standard setting. In practice this means that students (with a given score) are excluded only if one of the two tasks in their assignment was very good and the other very bad. This is the Achilles heel of the Body of Work method, and can lead to serious biases if the 'random picking' of the scripts or recordings is done on sight by someone involved in the construction of the test or decision making using the results of the test. A completely blind procedure (after exclusion of works that are too heterogeneous) similar to a coin tossing procedure is the only way to avoid biased selection.

Apart from the validity and practical requirements as discussed above, a number of principles of a more general methodological nature should be followed to make the statistical results stable and to avoid biases:

1. Each judge should have an equal number of set A and set B performances: eight of each in Speaking and twelve of each in Writing.

2. Each judge must receive performances with a wide range of scores; i.e., not all low scores or all high scores or all scores in the middle of the range.
3. The folder of scripts or the CD of spoken performances should not be presented to the judges in a haphazard way. Instead two principles are followed:
    a. Each judge rates a sequence of four homogeneous blocks of performances, i.e., all A or all B assignments. For example, a judge may start in Writing with 6 performances from set A, then 6 from set B, then again six As and ending with six Bs. (In Speaking the blocks contain 4 performances).
    b. The order of presentation of the performances is random, meaning that high and low scores can (and will) occur in each block, such that there is no systematic trend in the scores in the way they are presented to the judges.
4. On top of this, it was intended to construct the design in such a way that each of the 13 judges would have at least half of the performances in common with at least one other judge[41].

All these requirements taken jointly are quite complex and it is not even known if it is at all possible to fulfil them all at the same time. We did not find a design that fulfilled all requirements. In Table 2, the frequencies of stuldents having a different number of judges are displayed. It turns out that about 69% of the students were rated by 4 judges, and the others by three or five judges (in equal amounts).

Table 2. Frequency distribution of number of judges per student.

Number of judges	Writing	Speaking
3	12	8
4	54	36
5	12	8
Total	78	52

The construction of a good and balanced design requires a bit of puzzling, and it is never sure whether the design can be implemented as it was set up. In this case, two exceptions occurred, a trivial one and a serious one. The trivial one was that for some scores there were less works with the required score available than prescribed (see Table 1 for an example). The solution in this case is easy: just choose a work (randomly!) with a neighbouring score. The other exception was more serious: shortly before the standard setting event had to take place, it became clear that the internal panel would consist of 18 members instead of the 13 planned earlier and which number was used to make the data collection design. Adaptation of the design from 13 to 18 judges is not a trivial task, and in general such adaptations should be avoided as much as possible. Fortunately, around the same time it became clear that the number of

[41] In retrospect this criterion was not necessary and led to a suboptimal design. Originally it was added to offer the possibility in later analyses to compute agreement between judges, at least for some pairs of them, but the logic of the body of work method does not assume a high or low agreement beyond the effect of the student score.

external judges would consists of 8 experts, such that in total the standard setting would be done by 18 + 8 = 26 judges, a miraculously beautiful multiple of 13 that was used in the set up of the design. So the design developed originally for 13 judges was simply doubled.

As an illustration the design for Speaking and some comments on its construction principles is discussed in appendix 3.

It should be stressed that the judges must not have any information whatsoever about the scores given to the works they are judging.

3. The statistical model

The statistical model used for analysing the data departs from two basic assumptions. The first one is that students functioning at the B2 level will obtain on average a higher score than students not yet at that level. The second assumption is that students who actually are at the B2 level or higher, will obtain (on average) more often a 'Yes' answer by the judges than students not yet at B2. The former assumption is an assumption on the validity of the test score, the latter on the validity of the judgements in the Body-of-Work method.

The **definition** of the cut-off score or performance standard is the score for which the probability of obtaining a 'Yes' judgment equals **0.50.**

Three sources affecting the probability of saying 'Yes' are distinguished: the first is the **score** of the student, the second is the **leniency** of the judge and the third is the **difficulty** of the assignment. This suggests a **regression** model, where the dependent variable is the answer 'Yes' or 'No' (coded as '1' and '0', respectively) and the independent variables, the predictors are the three aforementioned sources: score of the student, leniency of the judge and difficulty of the assignment. Notice that the expected value (or average) of the answers is the probability of obtaining a '1'.

A problem arises because the probability of the answers is bound by zero and one, and a linear regression would predict (for some values of the predictor) probabilities outside the (0.1) interval. Therefore, the probability of the answer is not taken as the dependent variable, but a function of this probability. Such a function is called a **link function**, and the most commonly used link function with probabilities is the **logit** function, defined as

$$\text{logit}(p) = \ln \frac{p}{1-p} \qquad (1)$$

where p denotes the probability (of saying 'Yes') and ln(.) is the logarithmic function. As p goes from zero to one (but excluding these two values), the function value goes from minus infinity to plus infinity. And in particular, **when $p = 0.5$, then logit(p) = 0**.

The linear regression where the logit is the expected value of the dependent variable is called **logistic regression.** In the present case it can be written explicitly as

$$\text{logit}[P(Y_{vj} = 1)] = \ell_j + b_1 s_v + b_2 set_v \quad (2)$$

where the used symbols have the following meaning:

- Y_{vj} is the (coded) response by judge j on the performance of student v: 1 for a 'yes' and 0 for a 'no';
- ℓ_j is the leniency of judge j (to be estimated);
- s_v is the score of student v;
- set_v denotes the assignment of student v: it takes the value 0 if the set is A, and 1 if the set is B;
- b_1 and b_2 are the regression coefficients (to be estimated).

A problem in the estimation of the parameters is the presence of the nuisance parameters ℓ_j. There are several approaches possible to get rid of these parameters[42]. We define

$$b_0 = \frac{1}{J}\sum_j \ell_j \quad (3)$$

i.e. J is the number of judges and b_0 is the average leniency of the judges, and we define θ_j as the deviation of the leniency of judge j to the average, i.e.

$$\theta_j = \ell_j - b_0 \quad (4)$$

In a first approach, we ignore the differences between the judges, and replace the logistic regression model (2) by a simplified one:

$$\text{logit}[P(Y_{vj} = 1)] = b_0 + b_1 s_v + b_2 set_v \quad (5)$$

In the analyses, the coefficients b_1 and b_2 and the intercept b_0 are estimated as well as their standard errors. The estimation method is maximum likelihood[43].

4. Finding the performance standards

Once all the regression parameters are known (or estimated), it is not difficult to find the performance standard. Remembering the relation between a probability, p and its logit transformation (see equation (1)), we find in particular the equivalence

$$p = 0.50 \;\Leftrightarrow\; \text{logit}(p) = 0\,.$$

[42] The one we have chosen is a bit unorthodox, but in discussions with the statisticians of the centre, it appeared that a more orthodox approach (a so-called mixture model) leads to essentially the same results. Our approach has the advantage that it is readily understandable for readers with a modest background in statistics.

[43] Logistic regression as a statistical method is implemented in many statistical packages like SPSS or SAS. For the body of work method as discussed in this chapter an ad hoc computer programme is available. It can be downloaded, together with a user's manual from the resource page of the EALTA website.

This means that we have to find a score s such that the right-hand side of equation (5) equals zero. It is easily verified that the solution is given by

$$s = \begin{cases} -\dfrac{b_0}{b_1} & \text{if } set = 0, \text{ i.e. the set is A} \\ -\dfrac{b_0 + b_2}{b_1} & \text{if } set = 1, \text{ i.e. the set is B} \end{cases}$$

If the coefficient b_2 is not significantly different from zero – and it was not in any of the analyses that were carried out – than one can make an estimate of a single cut score which is given by[44,45]

$$s = -\frac{b_0 + 0.5 \times b_2}{b_1}.$$

For both skills, three analyses were run: one using only the data from the 8 external experts, one only using the data from the Saint Petersburg panel (18 judges) and one based on all data simultaneously. As the internal panel outnumbers the set of experts by more than a factor 2, the results of the joint analysis will be more similar to the panel results than to the experts' results, but from this it does not follow automatically what the most rational choice is to come to a final decision. This problem will come more to the foreground when the differences between judges are investigated.

Apart from this, the performance standards are also reported for the two assignments (or sets) separately and jointly.

The results for Writing are given in Table 3 and for Speaking in Table 4.

Table 3. Performance standards for Writing.

	Experts		panel		exp. & panel	
	value	SE	value	SE	value	SE
based on set A	**22.07**	0.772	**24.88**	0.462	**24.06**	0.405
based on set B	**22.99**	0.759	**25.28**	0.471	**24.60**	0.408
based on both sets	**22.53**	0.543	**25.08**	0.330	**24.33**	0.287

From Table 3, two interesting observations follow. The first is that for the experts as well as for the local panel the standard (value) is set a bit higher for assignment B than for A. Although the difference is not statistically significant, it is worthwhile to notice

[44] The correct procedure would be to redo the analyses and leave out the set predictor altogether. But the result will be very close to the one presented here. The 0.5 in the formula derives from the fact that there was an equal amount of set A and set B performances.

[45] The standard error of the cut-score is computed using the so-called delta method. This is a quite involved technique which is not discussed in this chapter. Standard errors of the performance standards are given in the ad hoc computer programme (see footnote 4).

that both panels (working independently from each other) came up with similar results. The second, and a much more important observation is the marked difference between experts and the local panel. The performance standard for the local panel is substantially higher than for the experts, so that it is not immediately clear what to choose.

Table 4. Performance standards for Speaking.

	experts		panel		exp. & panel	
	Value	SE	value	SE	value	SE
based on set A	**13.86**	0.628	**14.25**	0.168	**14.12**	0.227
based on set B	**14.67**	0.602	**14.72**	0.161	**14.72**	0.220
based on both sets	**14.26**	0.440	**14.49**	0.117	**14.42**	0.158

The observations from Table 4 are, first, that again the standard for set B is higher than for set A, but not significantly so. The other observation is that, as far as the performance standard is concerned, the local panel and the experts come to the same conclusion: the pair of cut-off scores is 14/15, 14 meaning that the B2-level has not been reached, and 15 indicating the minimal score that grants the B2 denomination.

Note that for both skills the standards are about halfway the originally selected score range as predicted by the Saint Petersburg colleagues.

5. The Cizek & Bunch approach

The technique of logistic regression may look cumbersome to scholars who are not used to statistical modelling. In their chapter on the Body-of-Work method, Cizek and Bunch use an estimation method which is less efficient than the maximum likelihood method, but which is intuitively more appealing. To make the statistical analyses as clear as possible for the panel members, this method has also been used. An example is given for the Speaking performances as judged by the internal panel.

In Table 5 a summary is given of the calculations to be done to have the approximate standard. In the leftmost column the scores of the selected students are listed. The second column, labelled fr(equency) gives the number of judgments that have been given to a performance with a score given in the same row. For example, it happened 29 times that a judge had to give a judgment on a performance of a student who got the score 11. The next column gives the number of ‘Yes’ responses.

Table 5. Summary of the judgments for Speaking.

score	fr.	#Yes	p	logit
11	29	1	0.050	-2.944
12	27	0	0.018	-4.007
13	20	1	0.071	-2.565
14	32	5	0.167	-1.609
15	40	32	0.793	1.341
16	30	29	0.952	2.979
17	18	18	0.974	3.611
18	36	36	0.986	4.290
19	24	24	0.980	3.892
20	32	32	0.985	4.174

The column labelled ‘p’ is meant to give the proportion of ‘Yes’ responses, but just taking the ratio would give zero as a result in row 2 and one in the last four rows, and for all these cases the logit transformation is not defined. Therefore, a small ‘correction’ is applied: the number of ‘Yes’ answers is increased by a half (in each row) and the frequency is increased by one, and the ratio is computed on these adapted numbers, such that the logit transformation is always possible. The results are given in the rightmost column.

In Figure 1 the scatter plot is given for the scores (horizontal axis) and the logit value of the ‘Yes’- proportions (vertical axis). The straight line is the linear regression line (regressing the logit on the score) and the dashed lines show graphically how the cut score is determined: starting at zero on the vertical axis, going to the line and then vertically downwards reaches the horizontal axis at the cut score (which in the example is 14.62, quite close to the final estimate of 14.49; see Table 4).

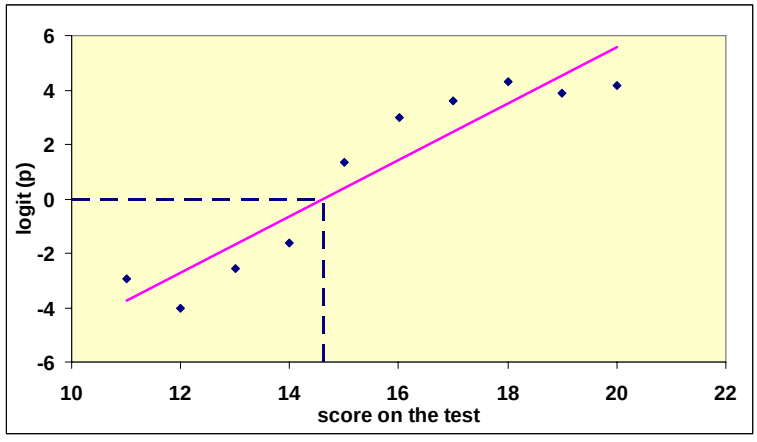


Figure 1. Approximate logistic regression for Speaking.

This last result does not mean, however, that the method shown in Cizek and Bunch is equivalent to logistic regression; it is not. In Table 6, the estimates of the three regression coefficients using the logistic regression techniques are displayed for Speaking and for the three regression analyses that were run.

Table 6. Regression coefficients for Speaking.

	experts		panel		exp. & panel	
	value	SE	value	SE	value	SE
b_0	-7.21	1.37	**-31.383**	4.94	-13.92	1.38
b_1	0.52	0.09	2.20	0.34	0.99	0.09
b_2	-0.42	0.45	-1.04	0.53	-0.60	0.31

There is one remark to be made concerning Table 6, viz. the high negative value for the intercept b_0 when using only the data from the local panel (bold faced in Table 6). The reason for this can be seen from Table 5 where there are five different scores having either not a single 'Yes' (one score:12), or not a single 'No' (the scores 17 to 20). This means that the panel is very homogeneous in the location of the cut score, and this also shows in the high value of the regression coefficient b_1 which is more than four times as high as for the experts: 2.202 vs. 0.521. This result shows also that one has to be careful with the Cizek & Bunch approach: the slope of the regression line, using this method (see Figure 1) is only 1.036, less than half the value of the maximum likelihood estimate displayed in Table 11. The reason for this is the effect of the so-called correction for the proportions, which is, although widespread, arbitrary and has a large

effect on the estimates of slope and intercept of the regression line. Fortunately, the effect on the estimate of the cut-score is very mild. In the logistic regression technique no such correction is necessary. But the risk of very high or very low proportions is the reason why the range finding round is so important when using the statistical technique proposed by Cizek and Bunch; using genuine logistic regression as we did does not have such a restriction.

From Figure 1 one can see that the individual points in the scatter diagram are relatively close to the regression line. This means that correlations between score and logit value are quite high. In this example it is 0.94. The correlations for all six analyses that were carried out are displayed in Table 7.

Although these correlations are a good validity argument – the standard setting judgments and the scores point to the same underlying construct – one should not be misled by their very high values, because they are partly an artefact of the way the student sample was selected. The sample is not a random sample from all participating students, but has been composed so as to obtain a (more or less) uniform distribution of the scores. Therefore the variance of the scores has increased (in comparison to a random sample), and this will automatically lead to an increase of the correlations. The interesting aspect in the table is the fact that the correlation for the panel is higher than for the experts, showing that the scoring system as used in Saint Petersburg and the judgement in the standard setting procedure are more homogeneous in the local panel than in the group of external experts.

Table 7. Correlations between logit values and scores.

	experts	panel	exp. & panel
Writing	0.88	0.95	0.95
Speaking	0.84	0.94	0.92

6. Differences between judges

6.1 Differences in leniency

As can be seen from equation (5), the index for the judges (*j*) appears at the left-hand side of the equation but not at the right-hand side, and this means that we have estimated the regression coefficients by treating all judges as 'equal'[46]. But of course, there may be (important) differences between judges and it is worthwhile to investigate them. Here we have taken an approach which brings the theory of the logistic regression in close relation to the Rasch model.

[46] In statistical jargon one would say: 'exchangeable'.

First let us use a shorthand notation for the right-hand side of equation (5) by defining

$$-\beta_v = b_0 + b_1 s_v + b_2 set_v . \tag{6}$$

So that we can rewrite (5) as

$$\text{logit}[P(Y_{vj} = 1)] = 0 - \beta_v \tag{7}$$

where the '0' has been added to indicate that the deviation from the average leniency (the intercept b_0) is zero for all judges. There exists a close relationship between a probability and its logit value defined by (1) but also by the inverse relation, finding the probability from its logit value. In the case of (7) this gives

$$P(Y_{vj} = 1) = \frac{\exp(0 - \beta_v)}{1 + \exp(0 - \beta_v)}, \tag{8}$$

in which one recognises immediately the logistic function. Using equations (2) and (4) we can write the model that takes differences between judges into account as

$$\text{logit}[P(Y_{vj} = 1)] = \theta_j - \beta_v , \tag{9}$$

and finding the probability itself from the logit we obtain

$$P(Y_{vj} = 1) = \frac{\exp(\theta_j - \beta_v)}{1 + \exp(\theta_j - \beta_v)} \tag{10}$$

which has the well-known appearance of the Rasch model, but in which the judges (indexed by *j*) play the role of 'persons' and the students play the role of 'items'[47]. The logistic regression with the simplified model has delivered already the value of $-\beta_v$ (which depends on the regression coefficients and the known score and set of the student; see equation (6)), so that we only have to estimate now the value of θ_j. In the present case we have used the Warm estimator, exactly as is done in the program package OPLM[48] (Verhelst, Glas & Verstralen, 1995; Verhelst & Glas, 1995).

Notice that positive values for this estimate indicate greater leniency than the average judge, while negative estimates indicate a harsher judge.

6.2 Differences in discrimination

The regression function allows to compute for every score the probability of a 'Yes' answer. From a judge who can discriminate well between low scores and high scores (of course without knowing them) we expect that he/she will say 'Yes' for students with a high probability and 'No' for students with a low probability. A measure, a kind of penalty, is the **residual sum of squares.** The residual is the difference between the

[47] In fact, the model we used in the logistic regression is equivalent to the Rasch model with linear restrictions on the item parameters (given by equation (6)). This model is also know as the Linear Logistic Test Model (LLTM; Scheiblechner, 1972, Fischer 1974).

[48] This estimator has been developed by Th. Warm (Warm, 1989). Its statistical merits are characterised by the fact that they are unbiased to a large degree. These estimates are also part of the ad hoc computer programme.

actual answer and the probability of a 'Yes'. This difference is squared to get rid of positive and negative numbers, and the squared residuals are then summed across all answers of the judge. In a formula this gives

$$RSS_j = \sum_{v(j)} \left[Y_{vj} - P(Y_{vj} = 1) \right]^2 \tag{11}$$

where the notation '$v(j)$' underneath the summation sign means that one takes the sum across all performances (students) that rater j judged. As the measure by itself is not very informative, we take a relative measure, which is the percentage of a rater's residual sum of squares with respect to the total residual sum of squares, labelled as the *percentage penalty* of judge j or PP_j:

$$PP_j = 100 \times \frac{RSS_j}{\sum_i RRS_i} \tag{12}$$

Notice that the smaller PP_j is, the better judge j discriminates.

Differences in leniency as well as in discrimination can be graphically displayed in one graph. Figure 2 displays the results for Writing and Figure 3 for Speaking. Each panel member is represented by a vertical line: a plain line for the Saint Petersburg panel and a dashed line for the external panel. The location of the lines corresponds to the θ-values of the judges and the heights of the lines represent the percentage penalty. As there are 26 judges in total, equal discrimination among judges would correspond to a PP_j value of 3.85.

For Writing it is remarkable that the external panel is as a whole more lenient than the Saint Petersburg panel. The average PP_j value for the external experts is 3.79 and for the local panel 3.87, a very small difference indeed. Of the local panel, 10 (out of 18) had a PP_j value larger than 3.85, and for only 2 out 8 external judges the PP_j value exceeded 3.85.

For Speaking the result is quite different: The PP_j value exceeded 3.85 for seven of the eight external judges and only for one member of the internal panel. Moreover, the internal panel is more homogeneous than the external panel: the standard deviation of their θ-values is 0.67, while in the external panel it is 1.93, almost three times as large.

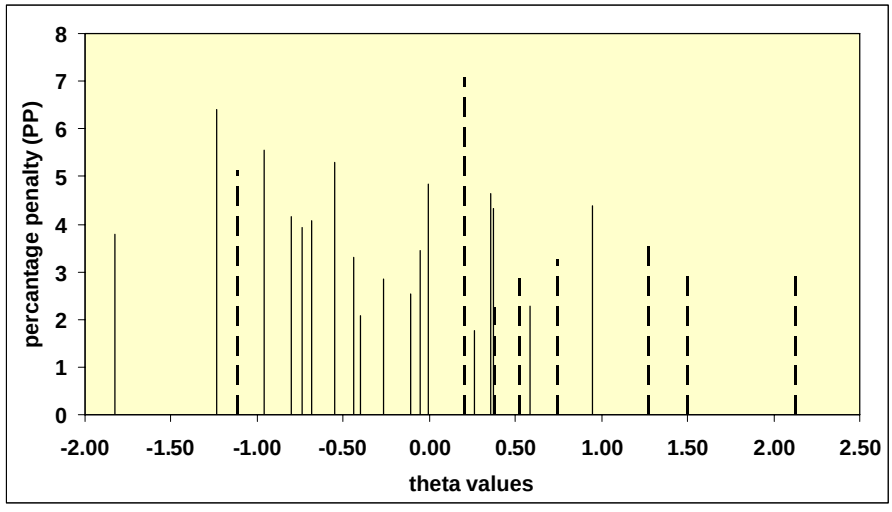


Figure 2. Leniency and discrimination for Writing.

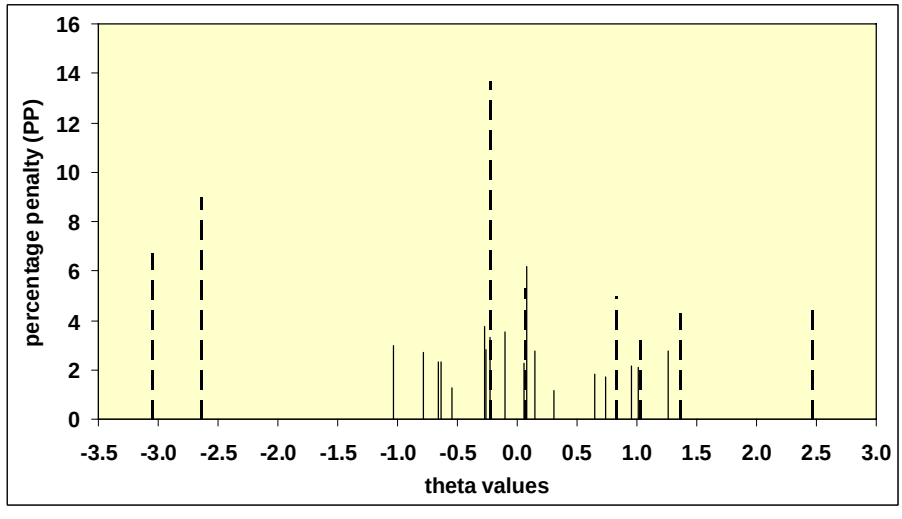


Figure 3. Leniency and discrimination for Speaking.

7. Conclusion

From an organisational and technical point of view the standard setting event was successful. Within three working days standards were set using 78 written and 52 spoken performances. One of the three working days was devoted to a rehearsal of the standardisation carried out in the months preceding the meeting and to presenting, explaining and discussing the results. The other two days were needed to do judgmental work and the initial workload of a maximum of 24 written performances and 16 spoken

performances per judge appeared to be very realistic: more work would undoubtedly have led to lessen the quality of the judgments. In a short questionnaire administered after the last judgment session, the 18 participants stated that they had understood the instructions, that they were confident about the accuracy of their own judgment[49], and that they would like to participate again in a similar panel.

Notice that panel members did judge 24 + 16 = 40 of the 130 performances used in the study, this is less than one third, and this was the reason that much attention was given to the design to collect the data. Sloppy designs can introduce all kinds of biases, and contra balancing possible effects makes the results more stable and trustworthy. Moreover, the use of an incomplete design made it possible to judge two different assignments in one event, giving the opportunity to have a least some view on the generalisability of the results.

The use of an incomplete design, however, also has a disadvantage: as each panel member had an (almost) unique subset of the performances to judge, organising a discussion round was practically impossible. So there was little opportunity to review one's views and convictions about the CEFR or about the requirements of the tasks and the quality of the performances, and this may be an important cause of the rather substantial variation in the leniency indicators (the θ-values).

The most puzzling outcomes are the systematic differences between the local panel and the external experts. For Writing the external experts are more lenient than the local panel and there is no clear cut explanation for this. For speaking there is a large amount of difference in leniency among the international experts and their judgments are far less consistent with the scores than the judgments of the internal panel. And also for this difference there is no clear cut explanation. The international experts did not form a panel (convening as a group at some point in time) and there was no specific training program for them: they were just asked for their expertise in Language Testing and for their familiarity with the CEFR. They came from three different countries (Sweden, Finland and Spain), and there was no independent check of their understanding of the CEFR or of their familiarity of the tasks which were developed by their Russian colleagues. Or more in general: it is maybe too early (or too naïve) to think one can give an answer to the question "Is your B2 also my B2?"

References

ALTE, s.d., Minimum standards for establishing quality profiles in ALTE examinations. *www.alte.org/resources/Documents/minimum_standards_en.pdf*.

Cizek G.J. & Bunch, M.B. (2007). *Standard Setting*. Thousand Oaks: Sage.

Council of Europe (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A Manual*. Strasbourg: Council of Europe. https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr

[49] One member of the internal panel stated that he/she was not confident for Speaking and two members were not confident for Writing.

Fischer, G.H. (1974). *Einführung in die Theorie Psychologischer Tests*. Bern: Huber

Kingston, N. M., Kahl, S. R., Sweeny, K. P. & Bay, L. (2001): Setting Performance Standards using the Body of Work Method. In Cizek G. J. (ed.), *Setting Performance Standards: Concepts, methods and perspectives*. Mahwah NJ: Erlbaum, pp. 219-248.

Kingston N.M. & Tiemann, G.C., (2012). Setting Performance Standards on Complex Assessments. In Cizek G.J. (ed.), *Setting Performance Standards: Foundations, Methods and Innovations* (second edition). New York: Routledge, pp. 201-223.

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben [The learning and solving of complex reasoning tasks]. *Zeitchrift für Experimentelle und Angewandte Psychologie, 3*, 456-506.

Verhelst, N.D., Glas C.A.W. & Verstralen, H.H.F.M. (1995). *One Parameter Logistic Model (OPLM)*. Arnhem: Cito

Verhelst, N.D. & Glas C.A.W. (1995). The One-Parameter Logistic Model. In G.H. Fischer and I.W. Molenaar (Eds), *Rasch Models: Foundations, Recent Developments and Applications*. New York: Springer-Verlag, pp. 215-237.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika, 54*, 427–450.

Appendix 1: Scoring rubrics for Spoken Interaction and Production Assessment Scale.

Band	Fluency and Content Relevance	Grammar: *appropriacy, control, range*	Lexical Resource: *appropriacy, control, range*	Pronunciation	Interaction
5	Fluent with **occasional** pauses and repetitions which are caused by the search of appropriate content or ideas. The response is **fully developed**; the content is **adequat**e to the task. Contributions are presented in a **logical sequence.** **Skillful use of various cohesive devices** to produce a coherent and logical utterance.	Produces **a mix of simple** and **complex structures** to complete the task. **Occasional** errors may occur in **complex structures**.	Can **handle without difficulty** a wide range of appropriate vocabulary. Occasional errors may occur in **less common lexical and idiomatic items and collocations.** **Can paraphrase effectively** to avoid repetition or achieve greater clarity or make language more expressive.	Rhythm and intonation are **generally appropriate**. Word stress is accurately placed. **May mispronounce** individual sounds in connected speech.	**Successfully** maintains communication towards an outcome: **- Keeps turn-taking** - Links contributions to those of the - partner to develop the communication - Can repair communication breakdown **Does not depend** on the partner.
4	*Some descriptors correspond to band «3», and some descriptors to band «5» or vary between bands «3» and «5»*				
3	Fluent, **though fluency can be disrupted at times** by pauses, repetitions or correction of mistakes. Corrects mistakes. The response is **generally full**; the content is **on the whole adequat**e to the task. **Logically** organised **ideas prevail.** Uses various cohesive devices.	Produces a mix of simple and complex **structures** to complete the task though the choice of structures **is repetitive.** **Errors occur mostly in complex structures.** **T**hose do not impede communication.	Has a **sufficient range** of vocabulary to complete the task though the word choice **may be repetitive**. Attempts to use less common vocabulary but not always successfully. Errors may occur, but they **do not impede communication**.	**May fail** to sustain the rhythm**.** May use a wrong intonation pattern. **Mispronounces** individual words and sounds in connected speech.	Can maintain communication **quite well** towards an outcome: - Keeps turn-taking - Considers partner's contributions - Can repair communication breakdown **Support** on the partner's part **may be required**.

2	*Some descriptors correspond to band «1», and some descriptors to band «3» or vary between bands «1» and «3»*				
1	**Fluency is disrupted.** Attempts to use complex structures result in **pauses and repetitions**. Complex structures may be left unfinished. The response may **fail to cover the key points** or **address the task**. Logical development and cohesion **are insufficient.** The range of cohesive devices is **limited.** Their use **may be inappropriate.**	Produces mostly **simple structures**. **Complex** structures **are attempted** but almost always contain errors. In these cases **some effort may be required to understand what has been said**.	Uses a range of **everyday vocabulary**. Errors may occur. **Some effort** may be required **to understand what has been said.**	**May often fail** to sustain the rhythm**. May often use** wrong intonation patterns. Makes **quite many errors** in pronunciation of individual words and sounds in connected speech.	Can maintain communication towards an outcome: - Keeps turn-taking at times - May fail to repair a communication breakdown **Support** on the partner's part **is required.**
0	*Performance does not satisfy the band «1» descriptors*				

Appendix 2: Scoring rubrics for Written Interaction and Production Assessment Scale.

	Content Relevance and Register Adequacy	**Vocabulary** *Appropriacy Range*	**Grammar** *Range Accuracy*	**Organisation and Mechanical Accuracy**
5	The response is **fully developed**; the content is **adequate** to the task. Contributions are presented in a **logical sequence.** Register is **appropriate.**	Can **handle without difficulty** a wide range of appropriate vocabulary. Occasional errors may occur in **less common lexical and idiomatic items and collocations. Can paraphrase effectively** to avoid repetition or achieve greater clarity or make language more expressive.	Produces **a mix of simple** and **complex structures** to complete the task. **Occasional** errors may occur in **complex** structures.	Layout **meets the requirements** of the task type. Uses paragraphs **correctly.** Punctuation and spelling are accurate, though **occasional slips** may occur. **Effective** use of a **variety of cohesive devices**.
4	*Some descriptors correspond to band «3», and some descriptors to band «5»*			
3	The response is **generally full**; the content is **on the whole adequat**e to the task. **Logically** organised **ideas prevail.** Register is **mostly** appropriate.	Has a **sufficient range** of vocabulary to complete the task though the word choice **may be repetitive**. Attempts to use less common vocabulary but not always successfully. Errors may occur, but they **do not impede comprehension**.	Produces a mix of simple and complex **structures** to complete the task though the choice of structures **is repetitive. Errors occur mostly in complex structures.** Those do not impede **comprehension**.	**On the whole** layout meets the requirements of the task type. **Errors in paragraphing** occur. **Occasional** punctuation and spelling errors **do not impede comprehension**. **Generally** uses cohesive devices appropriately.
2	*Some descriptors correspond to band «1», and some descriptors to band «3»*			

1	The response **only partly covers the key points** or **addresses the task**. Logical development is **not always sufficient;** some ideas can be left unfinished, repetitions may occur and oddities may happen.	Uses a range of **everyday vocabulary**. Errors may occur. **Some effort may be required to understand the text**.	Produces **mostly simple** structures. **Complex** structures **are attempted** but nearly always contain errors. **Some effort** may be required **to understand the text**.	**Some errors** occur in the layout and paragraphing. Spelling and punctuation errors occur quite often. Some of them **may impede** comprehension or **some effort may be required to understand the text**. The range of cohesive devices is **limited.** Their use **may be inappropriate.**
0	*Performance does not satisfy the band «1» descriptor*			

Appendix 3: The design for Speaking.

The design has been prepared on an EXCEL spreadsheet. In the next Figure the design for Speaking for assignment A is displayed.

score	judge 1	judge 2	judge 3	judge 4	judge 5	judge 6	judge 7	judge 8	judge 9	judge 10	judge 11	judge 12	judge 13
11	3	14								5	10		12
11						6	11		2			13	
12		7		8		13		14					
12	10		1		12				12				
12							4	5		15		8	
13			11	15	1						2		
13	1	13				8						15	
14							10		4				1
14				5		15							10
14							2	16		8	11		
15	12		4		9				10				
15		5		13	4			8					4
15							9		1	14		6	
16	4	16	12								3		
16				7	10	7			11				
16										6	9		11
17						16	1	15				14	
17	11	6	3					6					
18				14	3					16		7	3
18			9				12		3		1		
18	2	15				5						16	
19				6		14				7	12		9
19			2				3	13			4		
20	9		10		11				9				
20										13		5	2
20		8		16	2			7					

The 26 rows represent the students whose work has been judged and in the leftmost column their score is displayed. The rows are sorted by increasing score. The columns represent 13 judges.

In the first stage the works are assigned to the judges, where three principles are taken into account:

1. For each judge exactly 8 works are selected; these are represented by the shaded cells.
2. Each work is assigned to 4 judges, and if this proves impossible (or one does not find such an allocation) make sure that equality is approximated as well as possible. Here a design with 4 judges per work was not found; the distribution of works across judges is given in Table 2.
3. Avoid making cliques of judges, a subset of judges that judge a subset of the works and leave all the other woks to the other judges (which then automatically will also form a clique).

For the assignments B the allocation of works is a literal copy of the above assignment.

In the second stage, the order of presentation of the works to the judges is decided. In the column for judge 1 we find the numbers 1 to 4 and the numbers 9 to 12. This means that this judge will start with four works of assignment A, then judge four B assignments (the numbers 5 to 8; not shown) and then again four works of assignment A, and finish by four B assignments. Notice that the numbers of each subset of four are well spread across the student scores and are assigned in random order, i.e. one must avoid presenting works of the same assignment in increasing or decreasing order of scores. The order of assignment of the B works must be done in an analogous way; not shown in the figure.

Notice also that seven of the thirteen judges start with assignment A works (the numbers 1, 3, 5…) and the others with assignment B works.

“Maailman kivoin kirja!” – Portfolio nuorten oppijoiden englannin kielen osaamisen kasvun dokumentoijana peruskoulun luokilla 1-3 kaksikielisessä opetuksessa

Taina Wewer
Eurooppa-koulu Luxembourg I, Opetushallitus
(The European School Luxembourg I, Finnish National Agency for Education)

Abstract

The best book in the world!” Portfolio as a means to document growth of the English language skills in bilingual education in grades 1-3 of the comprehensive school. The language assessment study originally published in the English language and presented in Finnish in this article was composed of two individual language portfolio experiments for young language learners. Hence, the practitioner action case study comprised of two different sets and foci of data, and it had a descriptive and developmental intention. Both experiments took place in primary grades 1–3 in a multicultural and multilingual university teacher training school within the educational frame of bilingual Finnish-English instruction. The first portfolio experiment concerned traditional English as a Foreign Language instruction in the third grade, whilst the second experiment was connected with bilingual education in grades 1–2. The focal point of the investigation was the informativity of the language portfolio as an indicator of young learners’ English language proficiency and its development.

The experiences and views of teachers (n=6), pupils (n=37) and their parents (n=35) were gathered in respect of portfolio work using questionnaires and interviews. The observations and results obtained were congruent in both cases regardless of portfolio focus and parallel with prior Finnish language portfolio experiment research reports. The experiences and opinions of the participants were very positive. Young students did not perceive the portfolio as an assessment method per se but rather as an opportunity to concretely showcase their language proficiency, even when very minuscule. Additionally, the possibility to apply language in new occasions and connect English proficiency with other skills was more appreciated by third graders. The versatility and student-centeredness of portfolio tasks was found crucial to cater diverse learners, especially boys. Parents in turn were particularly content to have the opportunity to peek into their children’s thoughts, attitudes towards language study and motivation to learn English which were revealed in the portfolio tasks. Parents were also able to form a better understanding on the bilingual content studied in school. The

vast majority of all participants were clearly in favour with using the language portfolio as a complementary assessment method amongst other methods.

1. Johdanto

Kielitaidon arviointia englanti vieraana kielenä eli EFL-kontekstissa (*English as a Foreign Language*) on tutkittu runsaasti. Sen sijaan tutkimus arvioinnista kaksikielisen opetuksen CLIL-viitekehyksessä (*Content and Language Integrated Learning*), jossa vieras kieli on sekä oppimisen kohde että väline, on Suomessa toistaiseksi jäänyt varsin vähäiseksi erityisesti nuorten kieltenoppijoiden kohdalla. Jonkin verran tutkimusta Suomessa on kuitenkin tehty (ks. esim. Wewer, 2013; Wewer, 2014). Ongelmana kentällä on ollut mm. se, että kielitaidon arviointia ei aina ole tapahtunut, koska CLIL-opettajat ovat voimakkaan immersioperinteen vuoksi nähneet vieraan kielen vain välineenä (Hüttner, Dalton-Puffer & Smit, 2013), jota ei sen vuoksi olisi tarpeen arvioida lainkaan, eikä opetussuunnitelmiakaan ole aina ollut saatavilla (Wewer, 2014). Arviointiaiheeseen on 2010-luvulla kiinnitetty huomiota enemmän myös eurooppalaisella asiantuntijakentällä (esim. Heine, 2015; Leal, 2016; Massler, 2011; Massler, Stotz & Queisser; Zafiri & Zouganeli 2017). Tässä artikkelissa pitäydytään kuitenkin vain suomalaisessa arviointiviitekehyksessä.

Vaikuttaa siltä, että kielitaidon arviointi on ollut CLIL-opetuksen Suomen villi länsi ja käytänteet kirjavia. Tätä kuvaa keväällä 2012 suoritettu kyselytutkimus (Wewer, 2014), jonka mukaan kaksikielisen opetuksen kontekstissa kielitaidon arviointi ja palautteen antaminen on Suomessa alakouluissa ollut ”epäsäännöllistä, satunnaista, epäsuoraa sekä ennemmin vaikutelmiin kuin näyttöön tai opetussuunnitelmaan perustuvaa” (s. iv). Kyseisessä tutkimuksessa selvisi, että opettajien (n=42) käyttämistä arviointimenetelmistä observointi, kirjalliset testit ja keskusteleminen olivat yleisimmät kielitaidon arviointi- tai palautteenantomenetelmät, ja noin neljäsosa opettajista ilmoitti, että ei kerää arviointitietoa systemaattisesti lainkaan, mikä oli selvästi tuolloin voimassa olleiden Perusopetuksen opetussuunnitelman perusteiden (POPS 2004) ohjausnormien vastainen käytäntö. Oppilaiden itsearviointia arviointimenetelmänä käytti vajaa puolet (18/42) opettajista; portfolio oli arviointimenetelmistä kaikkein vähiten käytetty. Suomessa kuitenkin valtakunnallisen, yhteisen arviointisysteemin puute ja opettajien laaja pedagoginen, menetelmällinen vapaus mahdollistaa kirjavat arviointikäytänteet.

Muutama tutkimuksen haastatteluosioon osallistunut opettaja kuvaili oman koulun systemaattista arviointi- ja palautekäytäntöä itsearviointilomakkeineen ja todistuskaavakkeineen, mutta heitä oli hyvin vähän. Saman tutkimuksen mukaan palautetta huoltajille heidän lastensa englannin kielitaidosta ja sen edistymisestä antoi toisinaan tai harvoin suurin osa opettajista (28/33), eli vain harva opettaja välitti huoltajille säännöllistä tietoa lastensa kielitaidon kehittymisestä. Huoltajista (n=97) lähes puolet (48%) oli sitä mieltä, että he eivät saa tarpeeksi tietoa lapsensa kielitaidosta ja sen edistymisestä, ja 76 % toivoi tulevaisuudessa saavansa siitä enemmän tietoa. Kolmas-, neljäs- ja viidesluokkalaisista oppilaista (n=109) vain 8 % koki saavansa koulussa tarpeeksi

palautetta omasta englannin osaamisestaan kaksikielisen opetuksen tunneilla. Heistä 63 % toivoi saavansa enemmän palautetta englannin kielitaidostaan. Tarve arviointi- ja palautemenetelmien kehittämiselle kaksikielisessä CLIL-opetuksessa oli siten ilmiselvän suuri.

2. Kaksikielisestä ja perinteisestä englannin opetuksesta

Kaksikielinen opetus tunnetaan Suomessa myös aikaisemmalla, hieman harhaanjohtavalla nimityksellään vieraskielinen opetus (POPS, 2004) ja kirjallisuudessa myös akronyymillä CLIL. Kaksikielinen opetus voidaan väljästi määritellä seuraavasti (esim. Coyle, Hood & Marsh, 2010; Wewer, 2014):

> *CLIL on kaksitahoinen opettamisen lähestymistapa, jossa vierasta kohdekieltä käytetään yhdessä koulun opetuskielen kanssa ennalta asetettujen tavoitteiden suuntaisesti sekä vieraan kielen että oppisisältöjen oppimiseen.*

Käytännössä muita kuin kieliaineita opetetaan kahdella eri kielellä, tässä tapauksessa suomeksi ja englanniksi siten, että englannin kielisyöte ja harjoitteleminen tapahtuu oppimistilanteissa erityisesti ohjeiden antamisena, toiminnan järjestämisenä, sosiaalisena kanssakäymisenä sekä oppiaineiden opettamisessa ja opiskelussa. Tarkoitus siis on, että opitaan vierasta kieltä (kieli kohteena) samalla, kun sitä käytetään esimerkiksi matematiikassa tai taitoaineissa opiskeluun (kieli välineenä).

Kielitaitotavoitteet voivat vaihdella. Kaksikielisestä opetuksesta ja sen tavoitteista onkin määrätty laajasti ja hieman eri painotuksin kansallisissa Perusopetuksen opetussuunnitelman perusteissa (POPS, 2004; POPS, 2014). Koska kaksikielistä opetusta toteutetaan Suomessa niin monin eri tavoin ja eri laajuuksissa (ks. esim. Kangasvieri, Miettinen, Palviainen, Saarinen & Ala-Vähälä, 2012), yhtenäisten kansallisten linjojen yksityiskohtaisempi määritteleminen olisi hyvin vaikeaa, ellei jopa mahdotonta. Tässä artikkelissa raportoitu portfoliotutkimus on toteutettu vanhan POPSin (2004) ollessa voimassa, mutta uuden, vuoden 2016 syksyllä voimaan tulleen POPSin (2014) ideologisessa vaikutuspiirissä, joten alla käsitellään lyhyesti molempien POPSien avainmääräyksiä kielitaitoon liittyen.

Vanhempi perustedokumentti tyytyi mainitsemaan yleisesti, että ”[k]eskeisenä tavoitteena on se, että oppilaat voivat saada vankemman kielitaidon kuin tavallisessa opetuksessa kielten opetukseen varatuilla tunneilla” (POPS, 2004, s. 272). Nykyisessä POPSissa (2014, s. 89) on omaksuttu laaja-alainen kielikasvatusnäkökulma kielenopetukseen kielestä ja metodista riippumatta, kuten alla oleva katkelma kaksikielisestä opetuksesta osoittaa:

> *Kaksikielisessä opetuksessa pyritään saavuttamaan hyvä ja monipuolinen kielitaito sekä koulun opetuskielessä että kohdekielessä. Kaksikielisen opetuksen pitkäntähtäimen tavoitteena on perustan luominen elinikäiselle kielten oppimiselle sekä kielten ja kulttuurien moninaisuuden arvostamiselle.*

Kummassakin POPSissa todetaan, että opetuksen järjestäjä määrittää paikallisesti, miten kaksikielistä CLIL-opetusta toteutetaan ja minkälaista kielitaitoa tavoitellaan.

Kaksikielisen opetuksen kielitaitotavoitteet pitäisi siten määritellä kunta- tai koulutason paikallisessa opetussuunnitelmassa tarkemmin. Näin ei kuitenkaan ole aina ollut (Wewer, 2014) huolimatta siitä, että POPS (2004, s. 272) on selkeästi edellyttänyt minimissään tavoitellun kielitaitotason määrittämistä kielitaidon neljässä eri perusosa-alueessa (kuullun ja luetun ymmärtäminen sekä puhuminen ja kirjoittaminen) ja kulttuurisissa taidoissa. Tämä on luonnollisesti muodostanut esteen kielitaidon arvioinnin toteutumiselle. Parhaillaan voimassa oleva POPS (2014, s. 90) määrää opetuksen järjestäjän paikallisesti päätettäväksi kohdekielen sisällöt ja tavoitteet samalla lailla kuin aikaisempikin, mutta mainitsee, että apuna voi käyttää Eurooppalaista kielten taitotasojen viitekehystä (EVK, 2003). Dokumentti huomauttaa myös, että opetuksessa pitää huomioida eri oppiaineiden erityislaatuisuus ja niiden kieli (s. 92). Tämä onkin huomattavin seikka, missä CLIL-opetus eroaa EFL-opetuksesta sen lisäksi, että EFL-opetuksessa kieli on enemmän kohde kuin opiskelun väline.

Englanti vieraana kielenä -opetuksen valtakunnalliset tavoitteet ja sisällöt ovat olleet hyvinkin selkeät ja huomattavasti yksityiskohtaisemmat kuin kaksikielisessä opetuksessa erityisesti POPS 2004 -dokumentissa. Englannin oppiaineessa opintojen tavoite on saavuttaa yleinen, erityisesti sosiaalinen kielitaito, joka on Suomessa määritelty Eurooppalaisen viitekehyksen kielitaitotaksonomian avulla (POPS, 2014, s. 245–248), kun taas kaksikielisessä opetuksessa kielen oppiminen nivoutuu läheisesti eri oppiaineisiin ja niiden sisältöjen opiskeluun, jolloin opittava kielitaito on enemmän akateemista ja tiedonalakohtaista (ks. esim. Gottlieb & Ernst-Slavit, 2014; Llinares, Morton & Whittaker, 2012). Käytännössä suuntautuneisuus sosiaaliseen yleiskieleen ja tiedonalakohtaiseen akateemiseen kieleen ei aina ole selkeä, eivätkä rekisterit täysin vastakkaisia, vaan ennemminkin osin päällekkäisiä. Niiden välinen ero on suhteellinen ja liukuva (Snow ja Uccelli 2009, s. 115) – varsinkin, kun nuoret oppilaat ovat vielä aloittelevia kielenoppijoita.

2.1 Nuorten oppijoiden kielitaidon arviointi

Nuorina oppijoina voidaan pitää noin perusopetuksen alakouluikäisiä eli 6–13 -vuotiaita lapsia (Hasselgreen, 2005). Oppimisen arvioinnin kivijalkoina perusopetuksessa Suomessa ovat Perusopetuslaki ja Perusopetusasetus sekä kulloinkin voimassa oleva valtakunnallinen Perusopetuksen opetussuunnitelman perusteet -asiakirja, joita on velvoittavana noudatettava. Kielitaidon arviointi kontekstista riippumatta noudattaa samoja perusperiaatteita kuin muukin oppimisen arviointi perusopetuksessa. Perusopetuslain (628/1998, 22 §) mukaan arvioinnin tarkoituksena on ohjata ja kannustaa oppimista ja kehittää oppijan edellytyksiä tarkastella omaa opiskeluaan itsearvioinnin avulla. On merkittävää, että oppimista ja itsearviointia halutaan edistää ja tukea lainsäädännöllä. Lisäksi sama lakipykälä edellyttää arviointimenetelmien monipuolisuutta. Perusopetusasetus (852/1998, 10 §) puolestaan edellyttää muun muassa, että oppijan

edistymisestä on annettava riittävästi tietoa oppijalle itselleen ja hänen huoltajalleen. Samaa totesi POPS 2004 -asiakirja arvioinnista kaksikielisessä opetuksessa, ja periaate pätee CLIL-opetukseen myös nykyäänkin. Ongelmallista on, että riittävyyttä ei ole mitenkään määritelty, ja se lieneekin hyvin subjektiivinen käsitys, kuten Wewerin (2014) tutkimus osoitti.

Myös uusi POPS (2014, s. 92) edellyttää kielitaidon arviointia kaksikielisessä opetuksessa:

> *Arvioinnin tulee antaa opettajalle, oppilaalle ja huoltajille monipuolisesti palautetta oppilaiden aineenhallinnan ja kielitaidon kehittymisestä suhteessa opetukselle määriteltyihin tavoitteisiin. Oppilaan kielitaidon kehittymistä molemmissa kielissä seurataan eri oppiaineissa sekä kaikkien opettajien yhteistyönä että oppilaiden itsearvioinnin ja vertaisarvioinnin avulla. Tässä voidaan hyödyntää esimerkiksi Eurooppalaista kielisalkkua. […] Arvioinnissa huomioidaan myös oppiainekohtaisen kielen kehittyminen niiden kielellisten tavoitteiden osalta, jotka on paikalliseen opetussuunnitelmaan kirjattu.*

Aikaisempaan POPSiin verrattuna suurimmat muutokset ovat monipuolisen arvioinnin ja palautteen korostaminen riittävyyden sijaan sekä kielen kehittymisen seuraaminen oppiaineittain. Kielitaidon tasoa ja sen kehittymistä peilataan edelleenkin opetuksen järjestäjän määrittelemiin kielitaitotavoitteisiin (POPS, 2014, s. 92). Eurooppalainen kielisalkku mainitaan nimeltä yhtenä arviointimenetelmänä, mutta muutoin ei menetelmiä tai arvioinnin käytänteitä avata kuin periaatteellisella tasolla. POPS (2014, s. 47) on lanseerannut käsitteen *oppimista tukeva arviointikulttuuri*, johon liitetään lukuisia positiivisia attribuutteja. Se on mm. jatkuvaa, oppimisen aikana tapahtuvaa eli formatiivista, rohkaisevaa ja yrittämään kannustavaa, kriteeripohjaista, läpinäkyvää, osallistavaa, vuorovaikutteista, ja siihen sisältyvät oikeudenmukainen ja monipuolinen arviointi. Tällaisia luonnehdintoja oli jossain määrin näkyvissä myös aikaisemmassa POPSissa (2004, ks. s. 262–263).

Nuorten oppijoiden arvioinnissa on vielä otettava huomioon omia erityispiirteitään, jotka sivuavat läheisesti nykyisen POPSin arviointiajattelua. Esimerkiksi Hasselgreenin (2005, s. 38) mukaan ihanteellisimmillaan nuorten oppijoiden arvioinnissa on pelillisiä elementtejä; arviointitehtävät ovat hauskoja, moniulotteisia, ikäkaudelle sopivia ja mielenkiintoisia. Lisäksi arviointitehtävien pitää olla informatiivisia arvioinnin kaikkia osapuolia eli oppijaa, huoltajaa ja opettajaa ajatellen. Pinter (2011, s. 35–36) lisää, että arviointitehtävien tulisi olla yksinkertaisia, konkreettisia ja ’tässä ja nyt’ -tyyppisiä; ne voivat sisältää ryhmätyöskentelyn elementtejä ja vertaisarviointia. Edelleen hän toteaa, että nuorten oppijoiden pitäisi saada tarvittaessa apua arviointitehtäviin, joiden pitäisi perustua lapsen aikaisempaan kokemusmaailmaan. Hänen mielestään arviointitehtävien pitää tukea muun muassa oppijoiden positiivisen kieliminän ja -itseluottamuksen kasvua sekä tietoisuutta omasta oppimisesta esimerkiksi itsearvioinnin keinoin. Toisin sanoen nuorien oppijoiden arviointi perustuu tuttuihin elementteihin – samantyyppisiin tehtäviin ja aktiviteetteihin kuin tunneilla muutoinkin tehdään. Suomessa tähän kuuluu lakitekstin mukaan myös itsearviointi.

Tehokas itsearviointi sisältää aina reflektointia (Alanen & Kajander, 2011) eli asioiden, kokemuksien, tapahtumien tai oppimisen pohtimista siten, että ne selkiytyvät omassa mielessä, ja niitä voi hyödyntää myöhemmässä oppimisessa (Vickery, 2014, s. 79). Alanen ja Kajander (2011) kuitenkin huomauttavat, että itsearviointi ja reflektointi eivät ole automaattisesti synonyymejä: itsearviointiin liittyy tavoitteiden tarkastelu, kun reflektio puolestaan edistää itsetuntemusta ja on tietoista pohdintaa omasta oppimisesta. Reflektio johtaa ymmärrykseen ja vastuun ottamiseen omasta oppimisesta (Vickery, 2014, s. 84). Nuoret oppijat voivat oppia reflektoimaan harjoittelemalla erilaisten tekniikoiden, esimerkiksi opettajan apukysymysten avulla (ks. esim. Costa & Callick, 2008; Vickery, 2014). Fernsten ja Fernsten (2005) korostavat, että reflektointi ja palautekatselmukset kuuluvat erottamattomasti juuri portfolioarviointiin, koska ne monen muun argumentin ohella tekevät oppijan metakognitiot eli ymmärryksen omasta osaamisesta näkyviksi, edistävät oppijan autonomiaa ja mahdollistavat jaetun, oppimista tarkastelevan diskurssin eri toimijoiden välillä.

2.2 Portfolio oppimisen ja arvioinnin tukena

Euroopassa käynnistyi 2000-luvulla Eurooppalaisen kielten viitekehyksen (CEFR, 2001; EVK, 2003) julkaisua seurannut kieliportfoliobuumi, sillä viitekehykseen liittyen oli kehitetty yhteiseurooppalainen kieliportfoliomalli. Eri EU-maille ja -kielille, eri-ikäisille oppijoille akkreditoituja *European Language Portfolio* (ELP) -kielisalkkuversioita oli mahdollista tarkastella ja vertailla Euroopan neuvoston sivustolla[50]. ELP-kielisalkussa on kolme osaa: 1) kielipassi, joka sisältää kieliprofiilin eli kielenkäyttäjän itsearvioinnit eri kieltensä taitotasoista perustuen EVK-viitekehystaksonomian kriteereihin, 2) kielibiografia eli kielenoppimiskertomus, jonka avulla kielenoppija voi avata omaa kielenoppimishistoriaansa ja -polkujaan, sekä 3) työkansio, joka sisältää reflektoituja, autenttisia ja tekijänsä itse valikoimia näytteitä kielellisestä osaamisesta.

Suomessa on kokeiltu ja tutkittu kielisalkkua eri kielissä myös alakouluikäisille oppijoille rohkaisevin tuloksin jo noin neljännesvuosisata sitten (ks. esim. Linnakylä, Pollari & Takala, 1994; Kohonen, 2005). Raporteista on käynyt ilmi monen muun seikan lisäksi, että kielisalkkutoiminta esimerkiksi korostaa oppimisen yksilöllisyyttä ja auttaa rakentamaan oppijoiden kieliminää (esim. Aula, 2005; Perho & Raijas, 2011) sekä metalingvistisiä taitoja (Viita-Leskelä, 2005). Haasteita puolestaan esiintyi muun muassa puhutun kielen taltiointiin liittyen ja erityisesti poikien haluttomuudesta tehdä kielisalkkuun liittyviä aktiviteettejä (Viita-Leskelä, 2005). Portfoliotyöskentelyä on luonnollisestikin kokeiltu myös muualla. On todettu, että kieliportfolio mahdollistaa oman persoonan, mielipiteiden ja ajatusten esille tuomisen toisella tavalla kuin perinteisessä arvioinnissa. Se kehittää oppijoiden reflektointitaitoja, ja ennen kaikkea port-

[50] Valtuutetut ja rekisteröidyt ELP-kielisalkut maittain englanniksi: https://www.coe.int/en/web/portfolio/accredited-and-registered-models-by-countrymodeles-accredites-ou-enregistres-par-pays (8.4.2018)

folio mahdollistaa vähäisenkin osaamisen esille tuomisen lasta kunnioittavalla ja motivoivalla tavalla (Ioanniou-Georgiou & Pavlou, 2003, s. 23; Jones 2012, s. 402, 414). Portfolion tarkoituksena on todentaa ja tehdä näkyväksi kielenopiskelijoiden oppimispolku sekä tavoitteiden suuntaisesti kumuloituva kielitaito, jolloin kielisalkkutyöskentely nähdään laajasti kokonaisuutena, jonka pedagogisia ja arviointiin liittyviä toimintoja ei ole tarpeen liikaa eriyttää (Hildén & Takala, 2005). Parhaimmillaan kieliportfolio dokumentoi sekä oppimisprosessin että lopputuotoksen, ja on konkreettinen osoitus oppilaiden moninaisista kyvyistä (Stefanakis, 2010, s. 10). Se on erityisen soveltuva oppimisen pitkittäistarkasteluun tavanomaisen lyhyen jakson läpileikkauksen (esim. perinteiset kokeet) sijasta (ks. myös Salo, Kalaja, Kara & Kähkönen, 2013).

Tässä raportoitavan kokeilun ja tutkimuksen tekemisen aikaan suomalaista kielisalkkumallia (EKS, 2014) ei ollut saatavilla. Lisäksi tutkimuksesta (Wewer, 2014) saatujen tulosten perusteella Suomessa kieliportfolio ei ollut laajalle levinnyt tahi suositttu arviointimetodi kaksikielisessä opetuksessa, jolloin oma portfoliomalli oli itse kehitettävä aikaisempien kokeilujen ja kirjallisuuden pohjalta. Esimerkiksi Smith ja Tillema (2003) erottavat neljä erilaista portfoliotyyppiä käyttötarkoituksen ja kontekstin mukaan: 1) työnäytteet (esim. referenssit), 2) harjoittelu (usein opetussuunnitelmaan ja oppimiseen liittyvä), 3) reflektio (itsearviointi, kasvu ja kehitys) ja 4) henkilökohtainen kehittyminen (esim. ammatti-identiteetti). Tutkija-opettaja päätyi oman mielenkiintonsa pohjalta ja Euroopassa näkemiensä erilaisten kielisalkkumallien innostamana kokeilemaan harjoittelutyyppistä portfoliomallia tietämättä, että Opetushallitus oli tilannut suomalaisilta yliopistoilta suomalaiseen peruskouluun oman kielisalkkumallin (EKS Tiivistelmä, 2014), joka oli samaan aikaan valmisteilla. Nykyään Eurooppalainen kielisalkku on Suomessa suositeltu arviointimetodi kaikessa kielikasvatuksessa kielestä riippumatta, myös kaksikielisessä opetuksessa (POPS 2014). Kielisalkku nähdään arviointimetodina, jossa on runsaasti uuden arviointikulttuurin piirteitä. Portfolio ylipäätään nähdään myös uusimmassa CLIL-kirjallisuudessa soveliaana kielitaidon arviointimenetelmänä (esim. Massler, Stotz & Queisser, 2014; Heine, 2015).

3. Portfoliokokeilut

Portfoliotutkimus toteutettiin alakoulussa kaksikielisen opetuksen kontekstissa (englantia vähintään 25 %) kahdessa eri osassa ja kahdella eri painotuksella: 1) englanti perinteisenä oppiaineena eli vieraana kielenä (EFL) ja 2) kaksikielinen opetus (CLIL). Portfoliokokeilut ankkuroituivat osaksi aikaisempaa kaksikielisen opetuksen arviointikäytänteiden tutkimusta Suomessa (Wewer, 2014). Kokeilujen tarkoituksena oli tutkia ja löytää tapoja hyödyntää kieliportfoliota alkuvaiheen kielenopiskelussa ja kielitaidon arvioinnissa sekä rikastaa sitä kautta formatiivisen arvioinnin menetelmällistä työkalupakkia. Toimintatutkimuksessa kokeillut portfoliot voidaan luokitella harjoitteluportfolioiksi (ks. Smith & Tillema, 2003), joissa oli elementtejä ELP-mallista. Portfoliokokeilut toteutettiin vuosina 2011–2014: EFL-portfolio marraskuusta 2011 toukokuuhun 2012 ja CLIL-portfolio elokuusta 2012 toukokuuhun 2014. Kokeiluportfolioi-

den rakenne oli hyvin vapaamuotoinen, joskin kumpikin sisälsi lapsen kokemusmaailmaan ja ikätasoon soveltuvan kielibiografian ja painottui työsalkkuun. Kielipassiosiota ei ollut lainkaan.

Koulukontekstissa nuorten oppijoiden kanssa oli helpointa koota fyysisesti konkreettinen kieliportfolio, johon oli kätevää lisätä uusia osia eli sivuja. Kolmasluokkalaisten EFL-portfolio oli kulmalukkokansio, 1–2-luokkalaisten CLIL-portfolio oli A4-kokoinen ruutuvihko, johon oli kiinnitetty erilliset portfoliokannet. Kieliportfoliotehtäviin käytettiin aikaa vaihtelevasti: EFL-portfolioon joka toinen viikko yksi englannin tunti, ja isommissa projektitöissä enemmänkin, sillä oppilailla oli kolme englanti vieraana kielenä -oppituntia viikossa CLIL-kielipainotteisuuden vuoksi. CLIL-portfoliotyöskentelyä oli alkuun harvakseltaan, sillä kielibiografian taltiointi oli aikaa vievää. Kokeilun loppuvaiheessa portfoliotyöskentely oli viikoittaista ja säännöllistä.

Kumpikin portfoliokokeilu alkoi kielibiografialla, joka oli suunniteltu Perhon ja Raijaksen (2011) sekä käsillä olleiden ELP-mallien pohjalta. Kielibiografian eli kielenoppimiskertomuksen tarkoitus oli kartoittaa ja dokumentoida monikielisten ja -kulttuuristen oppilaiden kielihistoriaa, -kokemuksia ja -taustaa. Kolmasluokkalaiset EFL-portfoliota kokoavat oppilaat kirjoittivat itse ’Minun kielenoppimiskertomukseni’ -biografiansa ohjaavien kysymyksien avulla esseemuotoon, kun osin luku- ja kirjoitustaidottomat koulutulokkaat vastasivat suullisesti ’Minun kielitaustani’ -haastattelukysymyksiin, joiden vastaukset opettajaopiskelijat kirjasivat ylös. Kielibiografiaan sisältyivät lasten kotikielet, itse arvioitu kielitaito, kohtaamiset eri kielten ja kulttuurien edustajien kanssa, omat huomiot eri kielten olemassaolosta ja toiveet siitä, mitä haluaisi oppia englanniksi tulevaisuudessa. Näitä toiveita toteutettiinkin myöhemmin opetuksessa.

Nuorten oppijoiden arviointitehtävien erityispiirteitä (Hasselgreen, 2005; Pinter, 2011) huomioitiin kummankin portfoliokokeilun työsalkkutehtävissä. Ne perustuivat oppitunneilla käytettyyn sosiaaliseen ja akateemisempaan kieleen. EFL-portfoliotyöt suunnitteli ja ohjasi tutkija-opettaja englannin aineenopettajana yksin, ja ne liittyivät pääasiallisesti ja läheisesti englannin oppitunneilla käytettyyn oppikirjasarjaan, SanomaPron Yippee3! (Kuja-Kyyny-Pajula, Pelto, Turpeinen & Westlake, 2009), ja sen aihepiireihin (esim. *Super Toy*, *Imaginary Family*), mutta myös paikallisen opetussuunnitelman aihepiireihin (esim. *Menu*) ja Kyproksella sijainneen ystävyysluokan kanssa tehtyihin kommunikatiivisiin ja kulttuurisiin tehtäviin (esim. *Γεια σου*, itsensä esittely ja *Christmas in Finland*). Lisäksi portfolioon sisältyi joitakin reflektio- ja itsearviointitehtäviä (esim. *Benefits of Studying English*, *Week Reports*), ja ainakin yksi tehtävä muokkautui brittivaihto-opiskelijan luokkavierailun pohjalta (*School Uniform*).

Alkuopetuksen CLIL-portfoliokokeilu oli kaksin verroin pidempi, jolloin näytetöitäkin syntyi lukumäärällisesti enemmän. Harjoittelukoulun toiminnan luonteesta johtuen vaihtuvat luokanopettajaopiskelijat suunnittelivat yhdessä tutkija-opettajan kanssa, mutta myös omatoimisesti monia portfoliotöitä erityisesti luokkakieleen viikkoteemoihin ja ympäristö- ja luonnontietoon liittyen. Suuri osa ensimmäisen vuoden tehtävistä dokumentoi kumuloituvaa puhuttua yleiskieltä (esim. värit, numerot, ohjeisiin reagoiminen, angloamerikkalaiseen kulttuuriin lukeutuvaa sanastoa ja perinteitä),

koska oppilaiden suomen kielen luku- ja kirjoitustaidon edistäminen oli tavoiteprioriteetti. Toisena vuonna kirjoitetun kielen osuus kasvoi, samoin tiedonalakohtaisen kielen karttumisen ja ymmärtämisen dokumentointi erityisesti matematiikassa (esim. laskutoimituksiin liittyviä termejä, geometristen muotojen nimitykset) ja ympäristö- ja luonnontiedossa (esim. planeettojen nimet, ilmansuunnat). Tarkemmat tehtäväkuvaukset voi lukea alkuperäisen raportin Liitteistä 1 ja 2 (Wewer, 2015).

3.1 Tutkimuskysymykset ja –menetelmät

Nuorten oppijoiden arviointitutkimuksessa voidaan erottaa sekä yleisiä että erityisiä tavoitteita (McKay, 2006, s. 65). Tämän tutkimuksen yleisenä tavoitteena oli muodostaa selkeämpi kuva kieliportfoliosta yhtenä nykyaikaisena, vaihtoehtoisena arviointitapana ja parantaa arvioinnin vaikuttavuutta arvioinnin eri osapuolten eli oppilaiden itsensä, heidän huoltajiensa sekä opettajan näkökulmasta. Lisäksi tavoitteena oli saada yleiskäsitys oppijoiden kielitaidosta ja sen kehittymisestä. Erityisinä tavoitteina voidaan pitää vastaamista tutkimuskysymyksiin, joita oli neljä. Ensimmäisen kysymyksen tarkoituksena oli selvittää, minkälaisena oppilaat ja heidän huoltajansa näkivät portfolion kielitaidon dokumentoijana ja kehityskaaren esille tuojana.

1) Kuinka informatiivisena oppilaat ja heidän vanhempansa pitävät kieliportfoliota kielitaidon ja sen kehittymisen indikaattorina sekä kaksikielisessä CLIL-opetuksessa että perinteisessä englannin opetuksessa?
   Toinen ja kolmas kysymys liittyivät portfolion ominaisuuksiin ja laatuseikkoihin arviointimenetelmänä.
2) Mitä mielipiteitä ja kokemuksia opettajilla, oppilailla ja heidän vanhemmillaan on kieliportfoliosta?
3) Mitkä ovat kieliportfolion edut ja haitat arviointimenetelmänä?

Viimeinen, tulevaisuussuuntautunut kysymys liittyi arviointimenetelmän muokkaamiseen tarpeita vastaavaksi. Tässä artikkelissa keskitytään vain kolmeen ensimmäiseen kysymykseen.

Koska laadullinen toimintatutkimus ei luonnollisestikaan voi olla täysin objektiivinen (Cohen, Manion & Morrison, 2007, s. 310), pyrittiin tässä lisäämään luotettavuutta paitsi osallistujien, myös menetelmien triangulaatiolla, jolloin näkökulmia tutkittavaan aiheeseen saatiin enemmän. Koehenkilöjoukko jaettiin kolmeen ryhmään: ensisijaisiin ja toissijaisiin osallistujiin sekä avustajiin. Ensisijaisia osallistujia olivat sekä noin 7–10 -vuotiaat vuosiluokkien 1–3 oppilaat (n= 37) yliopiston harjoittelukoulussa että heidän huoltajansa (n=35). Toissijaisia osallistujia olivat tutkija-opettaja itse, portfoliokokeilun aloittaneet opettajakollegat (n=2) ja portfoliotyössä avustaneet opettajaopiskelijat, joista haastatteluun osallistui kolme. Avustajiksi luokiteltiin sellaiset henkilöt, jotka satunnaisesti olivat vaikuttamassa portfoliotehtävien syntyyn ja kertymiseen. Tällaisia henkilöitä olivat esimerkiksi vierailijat, kyproslaisen ystävyyskoulun opettaja sekä natiiviopettaja, jonka keskustelutuokioissa syntyi suulliseen tuottamiseen

ja kuullun ymmärtämiseen liittyviä tehtäviä, jotka liitettiin mukaan portfolioon kummassakin kokeilussa.

Molemmissa tapaustutkimuksissa aineisto kerättiin puolistrukturoiduin kyselyin sekä vapaaehtoisia haastattelemalla. EFL-portfolioaineisto koostui 18 oppilaskyselystä ja 17 huoltajakyselystä; CLIL-portfoliokyselyyn vastasi 19 oppilasta ja 18 huoltajaa. Huoltajat saivat kieliportfolion kotiin nähtäväksi kyselylomakkeen täyttämisen avuksi. Tiettävästi tämä oli Suomessa ensimmäisiin lukeutuva, ellei jopa ensimmäinen kieliportfoliotutkimus, johon on otettu mukaan myös huoltajien näkökulma. Molempiin aineistoihin kuului myös seitsemän oppilashaastattelu-tallennetta, jotka litteroitiin minimitasolla siten, että täytesanat, väärät aloitukset yms. jätettiin pois. Oppilaat osallistuivat haastatteluihin vapaaehtoisina, ja nauhoitetussa teemapohjaisessa keskustelussa oli portfolio mukana muistin ja havaintojen tukena. Lisäksi CLIL-portfolion yhteydessä toteutettiin samalla periaatteella temaattinen, taltioitu ryhmähaastattelu, johon osallistuivat alkuvaiheen työskentelyssä mukana olleet opettajaopiskelijat (n=opettajaedustajina. EFL-portfoliokokeiluun ei osallistunut opettajaopiskelijoita.

Lisäksi CLIL-kokeilun aloitti kaksi opettajakollegaa, toinen kutsuttuna, toinen itse kiinnostuksensa osoittaneena vapaaehtoisena. Kumpikaan ei saanut toimintaa ankkuroitua luokkansa arkeen, joten kokeilut eivät edenneet alkuaan pidemmälle. Kutsuttu opettaja kertoi myöhemmin osallistuneensa luokkansa kanssa myös toiseen kokeiluun, joka tuntui opettajasta mielekkäämmältä, ja portfoliotyöskentely jäi oppilaiden oman mielenkiinnon ja aktiivisuuden varaan oppituntien lisätyöksi. Vapaaehtoisen opettajan syyt johtuivat todennäköisesti portfolion työllistävyydestä ja työnkuvasta harjoittelukoulun opettajana, jolloin portfolion aloittaminen ja vakiinnuttaminen pitää ottaa osaksi harjoitteluluokan opetussuunnitelmaa ja sen ylläpitäminen vaatii vaihtuvien opettajaopiskelijoiden perehdyttämistä ja sitouttamista systemaattisesti.

Aineiston purkaminen tapahtui sisältöanalyysin keinoin sekä laskien frekvenssejä ja prosenttiosuuksia, ja kummastakin portfolioaineistosta tehtiin erillinen, aihealueisiin perustuva analyysi. Aihealueet olivat osin samoja. EFL-portfolion aihealueet olivat ”mielipiteet portfoliotehtävistä, kielitaidon osoittaminen, kielibiografia, portfolio kielitaidon ja sen kehittymisen indikaattorina sekä tulevaisuuden näkymiä”, ja CLIL-portfolion aihealueet ”kieliportfolion kokeminen tärkeäksi, mielipiteet portfoliotehtävistä, kielibiografia, portfolio kielitaidon ja sen kehittymisen indikaattorina, portfolion ainespesifisyys ja tulevaisuuden näkymiä” (Wewer, 2015, Liite 7). Tuloksista luokitelluista aihealueista tunnistettiin yleisimpiä, toistuvia teemoja, ja niitä havainnollistettiin haastatteluista ja kyselylomakkeista poimittujen autenttisten esimerkkien avulla.

4. Tärkeimmät tutkimustulokset

Kävi selkeästi ilmi, että sekä opettajien, oppilaiden että heidän huoltajiensa mielipiteet ja kokemukset olivat erittäin samansuuntaisia kummassakin kieliportfoliokokeilussa huolimatta hieman erilaisesta painotuksesta (sosiaalinen vs. akateeminen kieli), joskin

CLIL-portfolio nähtiin hieman positiivisemmin kuin EFL-portfolio. Tämä johtui todennäköisesti siitä, CLIL-kokeilu kesti kauemmin kuin EFL-kokeilu, ja siinä näkyi oppilaiden kielitaidon kehittyminen huomattavasti voimakkaammin kuin EFL-portfoliossa. Molemmat koeryhmät olivat kaksikielisessä opetuksessa vielä hyvin alkuvaiheessa, jolloin sosiaalinen ja akateeminen kieli eivät olleet voimakkaasti eriytyneet. Lisäksi CLIL-ryhmän opiskelumotivaatio ja luokan yleinen henki oli positiivisempi kuin EFL-ryhmässä, mikä osin näkyi oppilaiden kyselyvastauksissa. Molemmille ryhmille arvioinnin työtapa oli uusi. Tulosten ja havaintojen samankaltaisuuden vuoksi seuraavassa nostetaan esille merkittävimpiä tuloksia yhteisesti ja kootusti esittelemättä tuloksia täysin erillisinä.

4.1 Pääosin positiivisia merkityksellisyyden kokemuksia

Osallistujat suhtautuivat kieliportfolioon yleisesti varsin positiivisesti, joskin jotkut osallistujat huomauttivat aiheellisesti, että suullisen kielitaidon dokumentointiin kirjallinen portfolio ei ole sovelias. Kolmasluokkalaisista EFL-oppilaista 89 % oli sitä mieltä, että portfolio on hyvä kielitaidon indikaattori, ”[k]oska portfoliotyö näyttää, mitä olet oppinut ja mitä opettelet” ja ”[s]iitä näkee että kuinka hyvin osaan ja että kuinka hyvin kirjoitan”. Samaa mieltä oli 82 % heidän huoltajistaan. Oppilaista 65 % totesi, että heidän englannin kielen tasostaan saa käsityksen portfoliota selailemalla; yhtä suuri osuus oppilaista huomasi myös edistymisestä kielitaidossaan kokeilukuukausien aikana.

Tutkimuskyselyjen aikaan toisluokkalaisista CLIL-portfolio-oppilaista 74 % ajatteli, että he voivat portfolion avulla osoittaa omaa englannin kielen taitoaan, kun 67 % oli sitä mieltä, että portfolio tosiasiassa edustaa heidän kielitaitoaan. Omia tuotoksia siinä olikin paljon vähemmän kuin EFL-portfoliossa, jossa oppilaat pystyivät osoittamaan myös kielellistä harrastuneisuuttaan ja osaamistaan kouluopintojen ulkopuolelta. CLIL-portfolio-oppilaiden huoltajista jopa 94 % piti kieliportfoliota hyvänä kielitaidon indikaattorina. CLIL-oppilaat pitivät omaa portfoliotaan itselleen joko erittäin tärkeänä (42 %) tai tärkeänä (53 %), ja he tuntuivatkin olevan hyvin ylpeitä omista kielellisistä saavutuksistaan. Joillekin portfolio toimi eräänlaisena kielimuistokirjana, toiset taas pitivät sen kokoamista vain yksinkertaisesti hauskana. Moni (7/19) huomautti, että portfoliossa on myös oppimisfunktio. Oppilashaastatteluissa kävi ilmi, että portfolio oli tärkeä siksi, että se edusti englannin kieltä, jota tarvitaan esimerkiksi matkustamisessa. He ajattelivat kieltä myös tulevaisuuden hyödyn kannalta.

Kummassakin ryhmässä oli poikaoppilaita (EFL-ryhmässä muutama, CLIL-ryhmässä yksi), jotka eivät nähneet kieliportfoliotyöskentelyssä mitään lisäarvoa, ”koska meidän pitää tehdä tylsiä tehtäviä”. Yksi poika sanoi vihaavansa portfoliota. Viita-Leskelä (2005) on tehnyt samanlaisia huomioita nimenomaan poikien suhtautumisesta kielisalkkutyöskentelyyn. EFL-luokan poikien kohdalla hypoteeseja negatiiviseen suhtautumiseen on useita. Mahdollisesti kyse oli esimurrosikäisten poikien yleisestä koulumotivaation puutteesta ja opiskelun vaikeudesta, joka puolestaan saattoi johtua siitä, että koulun opiskelukieli oli monelle toinen tai kolmaskin kieli. Pojilla oli

myös usein tapana harjoittaa koulutyössä välttely- ja vähättelytaktiikkaa ja siten alisuoriutua; vaikeana koetut sisällöt ja asiat oli helpompi ohittaa osoittamalla välinpitämättömyyttä portfoliota kohtaan, sillä erityisesti EFL-portfolio oli luonteeltaan soveltava ja vaati kielisisältöjen hallintaa. Myös kasvojen menettäminen ryhmäpaineen alla saattoi vaikuttaa vastauksiin, vaikka kyselyt tehtiinkin yksin. Motivaation puute heijastui haluttomien poikien kyselyvastauksissa ja osin myös kielibiografiassa keksittyinä ja liioiteltuina lausuntoina, mitkä kirvoittivatkin hämmästyneitä, pettyneitä ja sarkastisiakin kommentteja heidän huoltajiltaan.

4.2 Kielibiografia: minä olen ainutlaatuinen kielenkäyttäjä ja -oppija

Tutkimuskyselyssä oppilaita pyydettiin lukemaan kokeilun alussa kirjoitettu kielenoppimiskertomus ja kirjoittamaan huomioitaan. Tämä reflektiota vaativa kysymys oli molemmille oppilasryhmille haastava – luultavasti siksi, että reflektointia ei ollut suoranaisesti harjoiteltu tarpeeksi, mutta myös oppilaiden nuoren iän vuoksi. Apukysymykset olisivat helpottaneet vastaamista. Oppilaiden suhtautuminen kielibiografiaan oli kaksitahoinen – osa näki siinä oman kielihistoriansa ja arvosti sitä; osa ei oikein tiennyt, mitä heiltä odotettiin, eivätkä he osanneet sanallistaa ajatuksiaan. Yksi EFL-portfoliokokeiluun osallistunut oppilas kommentoi esimerkiksi seuraavasti:

> *Huomaan että olen innokas oppimaan kieliä ja että pidän englannista ja että haluan oppia lisää.*

Yksi monikielinen, maahanmuuttajataustainen oppilas puolestaan huomasi, että hänen pitäisi panostaa enemmän äidinkieleensä. Maahanmuuttajataustaisten oppilaiden äidinkieli pitäisi nähdä voimavarana koulussa, mikä auttaisi lapsia itseäänkin arvostamaan omaa kieli- ja kulttuuritaustaansa. Kielten moninaisuutta ei vielä osata hyödyntää tehokkaasti kouluissa ja opetuksessa, vaan keskitytään ennemminkin kielten vertailuun (Pitkänen-Huhta & Mäntylä, 2014).

Vanhemmista enemmistö (53 %) kommentoi lapsensa kielenoppimismotivaatiota ja opitun kielen määrää positiivisesti vastatessaan kysymykseen tekstistä heränneistä asioista jopa hymiöiden kera, kuten alla olevassa esimerkissä.

> *Positiivisia ajatuksia, kertomus oli mukava, koska siinä huomasi, että lapsi on myös omasta mielestään oppinut paljon uusia asioita* ☺.

Jotkut huoltajat taas havaitsivat arvostuksen ja motivaation puutetta kielen oppimiseen, mikä voikin olla todellinen ongelma, jos pieni lapsi sijoitetaan kaksikieliseen opetukseen huoltajan päätöksellä lasta kuulematta. Myös CLIL-portfoliotyöskentelyyn osallistuneiden lasten huoltajat kommentoivat kielenoppimiskertomuksia varsin positiivisessa sävyssä, kuten alla olevassa esimerkissä (oppilaan nimi poistettu):

> *[Oppilaalla] on konkreettisia kokemuksia ja muistoja siitä kertonut mitä hän muistaa omasta lähimenneisyydestä. [Oppilas] on miettinyt tarpeellista kielitaitoa ja sen perusteella missä on viimeksi lomareissulla matkustettu. Nykyään [oppilas] pohtii kielten tarpeellisuutta laajemmin (ja haluaa oppia mahdollisimman monta kieltä).*

CLIL-portfoliota koonneiden oppilaiden biografian olivat kirjoittaneet opettajaopiskelijat haastattelun perusteella. Teemahaastatteluun osallistuneet opiskelijat pitivät oppilashaastattelutilannetta hyvin informatiivisena ja mielenkiintoisena, koska se mahdollisti tutustumisen oppilaaseen syvällisemmin, kun luokkatilanteissa ujot oppilaat uskalsivat avautua vapaammin. Oppilaantuntemus kasvoi. Oppilaille puolestaan heräsi tunnepitoisiakin ajatuksia heidän lukiessaan omaan kielibiografiaansa kirjattuja asioita:

> *Minulle tuli hassuja tunteita. Kun olin pienempi, en ollut yhtä rohkea.*
> *Silloin osasin vain vähän englantia, se tuntuu oudolta.*

Toisluokkalaisten oppilaiden lyhyet kommentit paljastavat reflektoimisen haastavuuden, mutta pienikin huomio on reflektion alku.

4.3 Laaja-alaisuutta ja moninaisuutta tehtäviin

Oppilailta kysyttiin, mistä portfoliotehtävistä he pitivät eniten, mistä vähiten. Kummassakin kokeilukyselyssä oppilaiden vastaukset hajaantuivat laajasti, ja käytännössä lähes kaikki tehtävät saivat mainintoja suosikkeina tai inhokkeina. Tämä osoittaa, että on erittäin tärkeää laatia mahdollisimman erilaisia ja vaihtelevia tehtäviä ja töitä, joita voidaan liittää kieliportfolioon, jotta heterogeeniset oppilaat ja heidän erilaiset mieltymyksensä voidaan ottaa huomioon. Jotta vähemmän motivoituneet oppilaat voidaan osallistaa portfolion kokoamiseen, olisi hyvä kysyä myös oppilailta tehtävä- ja aihepiirideoita. EFL-portfoliossa pidetyin tehtävä oli *Imaginary Family* (myös *My Family*), jossa leikattiin aikakauslehdistä itselle perheenjäsenet ja kirjoitettiin heistä kuvaus yksikön kolmannessa persoonassa. Oppilaita miellyttivät oikeaan elämään liittyvät kommunikatiiviset tehtävät, joissa sai käyttää mielikuvitusta ja luovuutta, ja jotka eivät olleet liian strukturoituja. Myös sellaiset tehtävät, joissa sai kertoa itsestään ja jotka sai tehdä yhdessä parin tai ryhmän kanssa, saivat kiitosta. Toisaalta taas jotkut oppilaat mainitsivat kirjoittamista tai piirtämistä vaativat tehtävät vähiten pidettyinä, mutta vielä useammin (33 %) itsearviointia tai reflektointia edellyttävät tehtävät nimettiin inhokkeina. Vanhemmista 35 % ei osannut valita yhtä tehtävää tai osiota muita kiinnostavammaksi, vaan he totesivat kaiken olevan mielenkiintoista tai kertoivat olleensa yllättyneitä käytetyn ja käsitellyn sanaston laajuudesta.

CLIL-portfoliota koonneet nuoremmat oppilaat keskittyivät tehtäväarvioissaan erityisesti kielelliseen helppouteen tai vaikeuteen. Osa kaipasi lisää haastetta, ja osa ei pitänyt jostakin ainespesifistä tehtävästä siksi, että kyseinen oppiaine ei ollut mieluinen, tai sen aihealue tai termistö oli ollut hankala. CLIL-portfoliokyselyssä yksikään tehtävä ei erottunut muita suositumpana, vaan jakauma oli hyvin tasainen. Vain

muutama tehtävä sai kaksi mainintaa (esim. *My Week* ja *Planets*), suurin osa yhden. Nuoremmille oppilaille itsestä kertominen ja oman osaamisen esille tuominen vaikutti olevan hieman tärkeämpää kuin kolmasluokkalaisille, joista erityisesti introvertimmat oppilaat pitivät muita enemmän omaan persoonaan liittyvistä tehtävistä. Oppilaiden huoltajat olivat eniten panneet merkille tehtäviä, joilla oli suora yhteys lapsen elämään ja jotka paljastivat heidän ajatuksiaan, näkemyksiään ja mielipiteitään. Vanhemmat saivat siten uuden kurkistusikkunan lapsiinsa kouluroolissaan. He myös ilmaisivat hämmästystä siitä, miten paljon lapset jo oppineet englantia kaksikielisessä opetuksessa.

On todettava, CLIL-portfoliotehtävien laatu sekä lingvistinen ja pedagoginen lähestymistapa portfoliotyöskentelyyn vaihteli runsaasti, koska portfolio oli osa luokan opetussuunnitelmaa ja siten osa luokanopettajaopiskelijoiden harjoittelua, joita oli kahden vuoden aikana toistakymmentä. Toisaalta, ilman opettajaopiskelijoiden panosta kieliportfoliota ei olisi harjoittelukouluympäristössä voinut toteuttaa lainkaan, koska he suunnittelivat ja opettivat valtaosan lukuvuoden tunneista. Opettajaopiskelijat saivat kokemusta ja yhden mallin portfoliotyöskentelystä, jonka he voivat valmistumisen jälkeen halutessaan lisätä omaan arviointityökalupakkiinsa.

4.4 Kieliminä, motivaatio ja systemaattinen ajankäyttö

CLIL-portfolion alkuun saattamisessa ja kielibiografiassa avustanut opiskelijaryhmä korosti erityisesti oppilaiden innokkuutta osoittaa omaa osaamistaan, oli se miten vähäistä tahansa, koulutaipaleen alussa:

> *Ja sitten kun siellä on se sivu missä on niitä että 'mitä osaan jo [englanniksi]', niin kun ne olivat niin innoissaan siitä että kun ne osaa jonkun ja sitten ne oli että "oota, oota, oota … Good morning!" ja sitten ne keksi sen ja niitten kasvot ihan loisti. Ja [oppilas] ei meinannut pysyä edes penkillä kun hän tiesi niin hyvin. Se varmaan motivoi just englannin tähän [oppimiseen] ja sitten ne saa varmaan itsekin käsitystä siitä mitä ne osaa jo.*

Motivaatiotekijät ja positiivisen kieliminän ja itseluottamuksen kehittyminen reflektion ja itsearvioinnin kautta ovatkin selkeitä kieliportfolion etuja, mikä tuli myös oppilaiden ja vanhempien vastauksissa esiin.

> *Kieliportfolio on hyvä väline lapsen kielitaidon kehittymisen seuraamiseen. Vanhemmat saavat arvokasta tietoa lapsen kielitaidoista. Lisäksi lapsi oppii arvioimaan omaa osaamistaan – taito, jota tarvitaan!*

Oppimismotivaatio tuntuu kasvavan, kun oppilaat saavat konkreettisesti nähdä oman kielitaitonsa kasvavan, jolloin vertailukohde on oma osaaminen, eikä luokkatoverien. Yksi oppilas kiteyttikin portfolion evidenssiperusteisuuden toteamalla, että portfolio on hyvä osoitus hänen kielitaidostaan, "koska siinä on melkein kaikki mitä osaan englanniksi".

Opettajalta puolestaan vaaditaan pitkäjänteisyyttä, suunnitelmallisuutta ja tavoitetietoisuutta, jotta portfoliotyöskentely onnistuu. Kieliportfolion menestys riippuu paljolti opettajan innostuksesta, innovatiivisuudesta ja sitkeydestä, kuten kävi ilmi portfoliotyön aloittaneiden, mutta pian lopettaneiden opettajakollegoiden tapauksesta. Suomessa opettajan pedagoginen vapaus ulottuu myös arvioinnin keinoihin ja menetelmiin, joten opettajalla on oikeus omaksua tai hylätä valitsemansa arviointikäytänteet, ellei sitten esimerkiksi koulun tasolla ole yhdessä sovittu yhtenäisistä käytänteistä. Portfoliolle onkin hyvä varata säännöllinen aikansa, jolloin tuotetaan näytekansioon lisämateriaalia, arvioidaan omia tuotoksia, reflektoidaan kielitaidon eri osa-alueita ja asetetaan uusia tavoitteita.

5. Pohdinta

Kieliportfoliotyöskentelyn olennainen osa on reflektio. Oman osaamisen pohdinta ja kokemuksien siirtäminen uusiin yhteyksiin syventää oppimisen ja osaamisen ymmärrystä, ja on yksi portti oppimaan oppimiseen. Kirjallisuudessa korostetaan sitä, että myös nuoret oppijat on mahdollista harjaannuttaa reflektoimaan omaa osaamistaan ja toimintaansa. Tämä vaatii aikaa ja suunnitelmallisuutta, mallintamista, esimerkkejä ja apukysymyksiä. Perusopetuslain kirjain ja oletus itsearvioinnista toteutuu erinomaisesti esimerkiksi reflektoivassa portfoliotyöskentelyssä, joka mahdollistaa myös oppijan syvemmän itsetuntemuksen ja tekee oppimisesta merkityksellisempää – kieltä opitaan itseä varten, osin omien valintojen kautta, ja samalla tietoisuus omasta itsestä kielenoppijana kasvaa.

Kielibiografian kirjoittaminen ja täydentäminen osoittaa konkreettisella tavalla sekä oppijalle itselleen että opettajalle sen, mistä oppijan kieli-identiteetti on rakentunut. Se ei välttämättä ole aina lapsille itselleen selvää, eivätkä he vielä välttämättä tarkastele maailmaa kieliperspektiivistä. Monikulttuurisuuden ja -kielisyyden lisääntymisestä Suomen kouluissa on mahdollista ammentaa voimavaroja ja aiheita niin globaali- kuin kielikasvatukseenkin. Erilaisuus ei ole enää vierasta, eikä ole olemassa yhtä ainoaa kielenoppijaprofiilia. Oman ja muiden kulttuurien sekä kielien kunnioittaminen voi alkaa omasta luokasta, kun oppilaiden taustoista puhutaan avoimesti, jolloin he itsekin ymmärtävät oman kielensä rikkautena. Kielibiografia on hyvä työkalu tähän etenkin yksilötasolla. Portfoliotyöskentelyssä oppijat kartuttavat myös itsetuntemusta ja oppivat arvostamaan omia töitään. Mielekkäät ja oikeaan elämään sidotut, oppilaan omia valinta- ja vaikutusmahdollisuuksia sisältävät tehtävät ja projektit, joista monilla oli todellinen kommunikatiivinen tarkoitus, tuntuvat olevan sellaisia, joita nuoret oppijat erityisesti arvostavat. Introvertimmat oppilaat pystyvät myös ilmaisemaan itseään ja ajatuksiaan portfolion kautta. Kun opettaja tuntee oppilaansa ja heidän taipumuksensa sekä oppimistyylinsä, ryhmälle kohdistettu tehtävien laadinta onnistunee helpommin.

Huolimatta siitä, että tässä sekä aikaisemmissa suomalaisissa, nuoria oppijoita koskevissa kielisalkkututkimuksissa on saatu hyvinkin rohkaisevia, portfolion käyt-

töön kannustavia tuloksia, ei kielisalkku yksinään ole riittävä arviointimenetelmä. Uuden POPS 2014 -dokumentin mukainen arviointikulttuuri edellyttää monipuolista, jatkuvaa palautteen antamista oppimisen aikana. Kieliportfolio pitkittäisleikkauksena oppijan kielitaidon kehittymisestä ja oppilaslähtöisenä arviointityökaluna on hyvä perustyökalu etenkin tiedonalojen kaksikielisessä opetuksessa, mutta myös mainio lisä perinteiseen arviointiin muussa kielten opetuksessa. Suullisen kielitaidon arviointia ja tallentamista olisi myös syytä pohtia, koska portfolio on usein paperimuodossa. Digitalisaatio tarjonnee ratkaisun tähän ongelmaan. Käytettävissä olevat resurssit, pääasiassa aika, eivät useinkaan valitettavasti suosi kieliportfoliota – siksi se pitäisikin ottaa säännölliseksi osaksi luokkarutiineja, kuten nuorten oppijoiden arvioinnissa suositellaankin, jolloin se kertyy kuin itsestään. Opettajalta tämä edellyttää uudenlaista arviointiajattelua, tahtotilaa ja suunnittelua. Kun kyseessä on oppilaan sanoin ”maailman kivoin kirja”, koska se kirja kertoo juuri oppijasta itsestään, hänen osaamisestaan ja sen kasvustaan, lienee metodi kaiken vaivan arvoinen.

Lähteet

Alanen, R. & Kajander, K. (2011). Reflektio ja itsearviointi. Opettajan mielistelyä vai kriittistä oman toiminnan arviointia? Teoksessa R. Hildén ja O.-P. Salo (toim.), *Kielikasvatus tänään ja huomenna. Opetussuunnitelmat, opettajankoulutus ja kielenopettajan arki.* Helsinki: WSOYpro.

Aula, T. (2005). Kielisalkku peruskoulun alaluokkien englannin kielen opetuksessa. Teoksessa V. Kohonen (toim.), *Eurooppalainen kielisalkku Suomessa* (s. 103—116). Helsinki: WSOY.

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

Cohen L., Manion L. & Morrison K. (2007). *Research Methods in Education* (6. painos). Lontoo: Routledge.

Costa, A. L. & Kallick, B. (2008). *Learning and Leading with Habits of Mind: 16 Essential Characteristics for Success.* Alexandria, VA: ASCD.

Coyle, D., Hood, P. & Marsh, D. (2010). *CLIL. Content and Language Integrated Learning.* Cambridge: Cambridge University Press.

EKS (2014). *Eurooppalainen kielisalkku*. (14.4.2018). Haettu osoitteesta http://kielisalkku.edu.fi/.

EKS Tiivistelmä (2014). *Perusopetuksen Eurooppalainen kielisalkku Suomessa*. (8.4.2018). Haettu osoitteesta http://kielisalkku.edu.fi/wp-content/uploads/2014/08/EKS_tiivistelma_2014.pdf.

EVK (2003). *Eurooppalainen viitekehys: kielten oppimisen, opettamisen ja arvioinnin yhteinen eurooppalainen viitekehys*. Euroopan Neuvosto.Helsinki: WSOY.

Fernsten L. & Fernsten J. (2005). Portfolio assessment and reflection: enhancing learning through effective practice. *Reflective Practice* 6(2): 303–309.

Gottlieb, M. & Ernst-Slavit, G. (2014). *Academic Language in Diverse Classrooms. Definitions and Contexts.* Thousand Oaks, CA: Corwin.

Hasselgreen, A. (2005). Assessing the Young Language Learners. *Language Testing* 22(3): 337–354.

Heine, L. (2015). Leistungsmessung. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 20(2): 21-24.

Hildén, R. & Takala, S. (2005). Kielisalkulla kohti selkeää ja monipuolista arviointia. Teoksessa V. Kohonen (toim.), *Eurooppalainen kielisalkku Suomessa* (s. 315–326). Helsinki: WSOY.

Hüttner, J., Dalton-Puffer, C. & Smit, U. (2013). The power of beliefs: lay theories and their influence on the implementation of CLIL programmes. *International Journal of Bilingual Education and Bilingualism* 16(3): 267–284.

Ioanniou-Georgiou, S. & Pavlou, P. (2003). *Assessing Young Learners*. Oxford: Oxford University Press.

Jones, J. (2012). Portfolios as 'learning companions' for children and a means to support and assess language learning in the primary school. *Education* 40(4): 401–416.

Kangasvieri, T., Miettinen, E., Palviainen H., Saarinen, T. & Ala-Vähälä, T. (2012). *Selvitys kotimaisten kielten kielikylpyopetuksen ja vieraskielisen opetuksen tilanteesta Suomessa: kuntatason tarkastelu*. Jyväskylän yliopisto: Soveltavan kielentutkimuksen keskus.

Kohonen V. (toim.) (2005). *Eurooppalainen kielisalkku Suomessa. Tutkimus- ja kehittämistyön taustaa ja tuloksia.* Helsinki: WSOY.

Kuja-Kyyny-Pajula, R., Pelto, P., Turpeinen E. & Westlake, P. (2009). *Yippee! 3 Reader*. Helsinki: WSOYPro.

Leal, J.P. (2016). Assessment in CLIL: Test Development at Content and Language for Teaching Natural Science in English as a Foreign Language. *Latin American Journal of Content and Language Integrated Learning* 9(2): 293–317.

Linnakylä P., Pollari P. & Takala S. (toim.). (1994). *Portfolio arvioinnin ja oppimisen tukena*. Jyväskylä: Jyväskylän yliopistopaino.

Llinares, A., Morton, T. & Whitaker, R. (2012). *The Roles of Language in CLIL*. Cambridge: Cambridge University Press.

Massler, U. (2011). Assessment in CLIL learning. Teoksessa S. Ioanniou-Georgiou & P. Pavlos (toim.), *Guidelines for CLIL Implementation in Primary and Pre-Primary Education* (s. 114–136).

Massler, U., Stotz, D. & Queisser, C. (2014). Assessment instruments for primary CLIL: the conceptualisation and evaluation of test tasks. *The Language Learning Journal* 42(2): 137–150.

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

Perho, K. & Raijas, M. 2011. Kielisalkkuprojekti ja venäjän kieli alaluokilla. Teoksessa R. Hildén & O-P. Salo (toim.), *Kielikasvatus tänään ja huomenna. Opetussuunnitelmat, opettajankoulutus ja kielenopettajan arki* (s. 183–205). Helsinki: WSOYpro.

Perusopetusasetus 852/1998. (8.4.2018). Haettu osoitteesta https://www.finlex.fi/fi/laki/ajantasa/1998/19980852.

Perusopetuslaki 628/1998. (8.4.2018). Haettu osoitteesta https://www.finlex.fi/fi/laki/ajantasa/1998/19980628.

Pinter, A. (2011). *Children Learning Second Languages*. Basingstroke: Palgrave Macmillan.

Pitkänen-Huhta, A. & Mäntylä, K. (2014). Maahanmuuttajat vieraan kielen oppijoina: monikielisen oppilaan kielirepertuaarin tunnistaminen ja hyödyntäminen vieraan kielen tunnilla. Teoksessa M. Mutta, P. Lintunen, I. Ivaska & P. Peltonen (toim.), *Tulevaisuuden kielenkäyttäjä – Language Users of Tomorrow.* AFinLAn vuosikirja 2014. Jyväskylä: Jyväskylän yliopistopaino.

POPS (2004). Perusopetuksen Opetussuunnitelman Perusteet 2004. (14.4.2018). Helsinki: Opetushallitus. Haettu osoitteesta www.oph.fi/download/139848_pops_web.pdf.

POPS (2014). Perusopetuksen Opetussuunnitelman Perusteet 2014. Määräykset ja ohjeet 2014:96. (14.4.2018). Helsinki: Opetushallitus. Haettu osoitteesta www.oph.fi/download/163777_perusopetuksen_opetussuunnitelman_perusteet_2014.pdf.

Salo, O-P., Kalaja, M., Kara, H. & Kähkönen, K. (2013). Kielisalkku kielikasvatuksen työvälineenä – Jyväskylän normaalikoulun kielisalkkuhankkeen taustoja ja tavoitteita. Teoksessa M. van der Berg, R. Mäkelä, H. Ruuska, K. Stenberg, A. Loukomies & R. Palmqvist (toim.), *Tutki, kokeile ja kehitä. Suomen harjoittelukoulujen julkaisu 2012* (s. 35–46). Helsinki: Yliopistopaino.

Smith, K. & Tillema, H. (2003). Clarifying different types of portfolio use. Assessment & Evaluation in Higher *Education* 28(6): 625–648.

Snow C. E. & Uccelli P. (2009). The challenge of academic language. Teoksessa D. R. Olson & N. Torrance (toim.) *The Cambridge Handbook of Literacy* (s. 112-133). New York: Cambridge University Press.

Stefanakis, E. H. (2010). *Differentiated Assessment. How to Assess the Learning Potential of Every Student.* San Fransisco, CA: Jossey-Bass.

Vickery, A. (2014). *Developing Active Learning in the Primary Classroom*. Los Angeles, CA: Sage.

Viita-Leskelä, U. (2005). Kielisalkkutyöskentelyä perusopetuksen alaluokilla (saksa, A2-englanti). Teoksessa V. Kohonen (toim.), *Eurooppalainen kielisalkku Suomessa* (s. 117–131). Helsinki: WSOY.

Wewer T. (2013). English language assessment in bilingual CLIL instruction at primary level in Finland: Quest for updated and valid assessment methods. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 18: 2, 76-87.

Wewer T. (2014). *Assessment of Young Learners' English Proficiency in Bilingual Content Instruction* CLIL, (väitöskirja, Turun yliopiston julkaisuja B-385). Turku: Turun yliopisto.

Wewer, T. (2015). *Portfolio as an Indicator of Young Learners' English Proficiency in Mainstream English Instruction (EFL) and Bilingual Content Instruction (CLIL),* (pro gradu-tutkielma, Turun yliopisto). Turku: Turun yliopisto.

Zafiri, M. & Zouganeli, K. (2017). Toward an Understanding of Content and Language Integrated Learning Assessment (CLILA) in Primary School Classes: A Case Study. *Research Papers in Language Teaching and Learning* 8(1): 88–109.

This volume implements the decision taken by the European Association for Language Testing and Assessment (EALTA) to honour the memory of Professor Sauli Takala, a founding member of the Association and its President from 2007 to 2010. The 23 entries by professionals from different countries and from different spheres are a reflection of his wide and diverse areas of expertise and interest, his many co-operations and networks, as well as his significant contribution to the language education and assessment community.

UNIVERSITY OF JYVÄSKYLÄ
CENTRE FOR APPLIED LANGUAGE STUDIES

EALTA
www.ealta.eu.org
EUROPEAN ASSOCIATION FOR LANGUAGE TESTING AND ASSESSMENT

ISBN 978-951-39-7748-1