

## Papers – Abstracts

### **Eli Moe, Lisbeth Brevik *CLIL – some teaching and testing results***

Politicians frequently suggest that the teaching of foreign languages is a disadvantage for the weaker students. We will put forward teaching and testing results that suggest the opposite.

CLIL refers to a method where a school subject is taught in a foreign language; pupils learn subject matter and language at the same time. This presentation refers to four schools that taught the social science curriculum in English (six 7<sup>th</sup> grade classes and one 10<sup>th</sup> grade class) in 2008/2009.

The presentation has its base in a project initiated by *The Norwegian Centre for Foreign Languages in Education*. In order to document effects of CLIL teaching, teachers and testers cooperated in developing two parallel language tests for the 7<sup>th</sup> grade and one test for the 10<sup>th</sup> grade.

Our findings include qualitative and quantitative data; pupils' and teachers' opinions, school leaving marks for the 10<sup>th</sup> grade, results from the two 7<sup>th</sup> grade tests – in the beginning and end of the school year, and results from a control group receiving "normal" English tuition.

Our quantitative data show that the language skills of CLIL students improve for weak as well as for strong students. In the control group, however, the standard deviation increases from test 1 to test 2; i.e. the gap between the strong and the weak students grows wider. The qualitative data confirm this.

CLIL teaching improves students' language skills. Strong and weak students benefit from more exposure to foreign language. This should be of interest to teachers, school leaders and educational politicians

### **Jamie Dunlea, Neus Figueras *Replicating a CEFR test comparison project across continents***

This paper will describe a collaborative project to validate the results of a test-centered standard setting workshop held in Japan by administering the test to EFL learners in Europe and employing a different, student-centered standard setting method to replicate the original results in a European context.

Replicating results with different standard setting methods is often suggested as one way of validating cut scores, but is seldom realized in practice due to practical considerations, limited resources, and the potential difficulty of interpreting differing results. In this case, the chance to replicate results from the original workshop is especially significant as it provides the opportunity to see if the interpretations of the CEFR by teachers in Japan, as reflected in the results of the original test-centered standard setting, will be replicated in a European context by using the Contrasting Groups student-centered method with teachers and learners who are potentially more familiar with the CEFR than their counterparts in Japan. The presentation will focus mainly on describing the scope of the collaboration and the research design for data collection. Although a detailed analysis of the data is not expected to be completed by the time of EALTA 2010, it is hoped that preliminary results will be available.

### **Alistair Van Moere *Usability testing with stakeholders in the development of a formative assessment of oral reading fluency***

This presentation describes the usability and validation of an automated assessment of oral reading fluency (ORF) in children aged 5 to 13 years. Oral reading fluency, defined as "the ability to read a text quickly, accurately, and with proper expression" (National Reading Panel, 2000), reflects the ability of readers to recognize whole words rapidly and effortlessly. Three measurable aspects of oral reading ability that contribute to ORF are reading rate (words correct per minute), reading accuracy (words correct/words attempted), and expressiveness (appropriate pausing, intonation, and phrasing). Together, these provide a barometer of a student's literacy, and can provide an early warning if the student begins to fall behind.

Traditional ORF assessment requires time-consuming individualized testing using "pencil and paper" in which a teacher follows along with a student who reads a passage. The development of an efficient formative assessment tool is therefore relevant. This presentation will demonstrate how speech processing technology presents a useful means for automating the scoring of oral reading fluency. Usability trials involving the test development team together with a local school district demonstrated how the assessment tool saved teacher time, gave feedback and encouragement to students, created a digital portfolio of students' reading performances, and could potentially involve parents in the process of reading together with their children. It will also report on the relation between ORF and language learning progress, in the context of an assessment taken annually by up to 4 million children in the US alone.

### **Brent Bridgeman *Fairness Issues in the Use of Automated Essay Scores Across Cultures***

Automated essay scoring is now used, in combination with human scores, in high stakes admissions tests such as the Graduate Record Examination (GRE) and the Test of English as a Foreign Language internet-based test (TOEFL iBT). Automated scores are developed so that, on average, human and machine scores will have the same mean and similar score distributions. However, this does not guarantee that human and machine scores will be the same in every subgroup. Indeed, there is evidence that the machine gives much higher (over 0.5 SDs higher) scores to GRE essays from mainland China. This appears to be a cultural rather than a language phenomenon because the machine advantage for essays from Taiwan is considerably smaller, and in other Asian language countries, such as Japan, there is essentially no difference between machine and human scores. Curiously, the machine advantage for essays from China appears to be much larger for GRE Issue essays than for TOEFL iBT Independent essays despite the similarity in these tasks and in the populations taking the two tests. In addition to summarizing the data from well over 200,000 examinees, the presentation will discuss some hypotheses to account for these apparently contradictory findings, and will also explain procedures put into place to minimize the impact of these human-machine discrepancies on reported scores.

### **Inmaculada Escudero *A collaboration model for Spanish as Second Language Certificate for immigrant workers***

Collaboration in language testing and assessment is considered to be as important as validity and reliability (EALTA, 2006). Language assessment programs are increasingly requiring researchers and specialists to work in collaborative groups instead of, or in addition to, requiring them to work isolated. This is especially important when we work with special targets, such as immigrant population. Theoretical and practical issues need to be taken into account in the design, use and interpretation of the results of such assessments. In this communication we present a new Spanish as Second Language Certificate for immigrant working population. The design processes of this certificate have followed three main steps: first, an analysis of linguistics needs of the target population; second, a rigorous construction phase of tasks and items; and third, statistical validation. All phases have been part of a collaborative approach model. This model conceives collaboration in two different levels: contexts and fields (schools and universities, government responsible politicians, testing organizations and teachers, test developers and researchers, testers and testees), and between colleagues with different specializations (comprehension processes researchers, teacher in classrooms, teacher educators, policy makers, test developers and researchers, and item writers and psychometricians).

### **Karin Vogt *Adaptations of CEF descriptors to local contexts***

The Common European Framework of Reference for Languages (CEF) calls on users to adapt its scales and descriptors to local contexts. In this context, we are seeing two contradictory developments. On the one hand, there is hardly any research on adaptations of CEF descriptors that would ensure the quality criteria Lenz and Schneider (2004) have stipulated for new or adapted descriptors (e.g. for the ELP). On the other hand, we have to state a proliferation of new descriptors that seem to come from nowhere. The purpose of the paper is to present the findings of a research project that has attempted to adapt scales and descriptors from the CEF, ensuring the quality standards mentioned by Lenz and Schneider (2004). In a methodology adapted from Smith & Kendall (1963), North (2000) and Kaftandjieva & Takala (2002), the adapted descriptors underwent a two-step qualitative validation procedure during which the co-operation with professionals and foreign language teachers was crucial. First, relevant target language use situations were identified with professionals and experts. As a second step, practitioners in VOLL were consulted in workshops and were asked to assign a CEF level to adapted descriptors, the results of which were CEF-based competence profiles for three occupations. The project did not only yield scales and descriptors for foreign language competence profiles for three professions, thus representing the local context. It also shed light on the way foreign language teachers understand and interpret CEF-based descriptors, and ultimately CEF scales themselves.

### **Cecilie Carlsen *Linking a computer learner corpus to the CEFR***

In this paper I will present a project which has been collaborative and multidisciplinary from the beginning to its present use, namely the linking of a computer learner corpus of Norwegian (ASK) to the CEFR. The project was a collaboration between Dr. Felianka Kaftandjieva and myself. In the present paper I will show how we collaborated to obtain a reliable linking of corpus texts to the CEFR, making ASK one of the few learner corpora available linked to the proficiency levels of the CEFR.

Using a computer learner corpus linked to the CEFR as our source of data, is fruitful for language testing, SLA-research and not to forget, for the CEFR itself. Several critics have mentioned the need of validating the level descriptors against empirical learner data. A corpus linked to the CEFR is particularly well-suited to this purpose. Taylor and Barker (2008) have pointed to the advantages of using computer learner corpora in language assessment. A learner corpus linked to the framework upon which the tests are based, is particularly useful since it allows us to investigate what learners can and cannot do at the levels at which the tests aim to measure. In SLA-research, levels of proficiency are often underestimated in studies of factors affecting language learning (Thomas 1996, Jarvis and Pavlenko 2008). The field of language assessment may contribute with insights on how to reliably place learner texts on different levels of proficiency. After introducing the linking project, I will show the results of a study based on ASK aiming at validating the CEFR level descriptors of coherence with particular focus on connectives (and, but, therefore, however etc.) in Norwegian. This latter project is part of the SLATE-network, a collaboration between language testers and SLA-researchers in Europe.

### **Glyn Jones, Kirsten Ackermann *Using CEF-derived guidelines in test development and assessment***

This paper will recount how two related resources were derived from CEF descriptors and then refined in collaboration with their intended users - item writers and markers – in the course of redeveloping a general English test.

The paper will firstly describe a set of guidelines developed to help item writers to construct test items whose difficulty level matches the intended CEF levels of the test.

Secondly, the paper will show how marking criteria based on the CEF help distinguish one level from another and allow us to decide whether a test taker is at the level he or she was entered for in a general English exam.

In both cases the approach adopted was to pinpoint key terms used in the CEF descriptors and to ask the question “What does this term mean in practice for this item type at this level?” The users of the respective documents, the item writers and marker, were involved in their development through practical trials and feedback questionnaires. The effectiveness of the marking criteria was further investigated using Rasch analysis of scores awarded in the course of trial administrations of the test.

The presenters will outline the process of developing the guidelines and criteria, with examples; discuss some of the problems of interpretation that arose in the process of formulating the guidelines and suggest some tentative solutions; present the findings of research into the usefulness of the guidelines as perceived by item writers and as evinced by the performance of raters.

### **Sahbi Hidri *Static vs. Dynamic Assessment of Listening Comprehension for Learners of English in an EFL Context: Two Complementary Approaches***

The testing of the language skills has been largely carried out in a traditional way where learners of English perform individually on a test. Recently, there has been a call for testing the language skills in a dynamic/collaborative way where more than one test taker can perform collaboratively on a test.

This paper compares and discusses static vs. dynamic assessment of listening comprehension (LC) for learners of English at the university level. The study collected qualitative and quantitative data: two LC tests, static and dynamic, for 60 tests takers, retrospective interviews for 60 test takers and 6 test raters. Results of the study indicated that the learners' LC ability is rated differently from one test mode to the other. Also, results of the study demonstrated that the same test items were scored differently in both test modes. Recommendations were made to highlight the integration of both assessment approaches to form a comprehensive view of the learners' LC ability in different contexts.

### **Henning Rossa *Validity inquiry on the process level: Think-aloud data on listening task performance***

The paper is concerned with the investigation of the construct validity of an EFL listening comprehension test that was developed for a large-scale assessment project. The study focuses on two objects of language testing research that share a paradoxical position in the scientific community: Both listening comprehension and validity are held in high esteem theoretically, but this has not motivated a significant number of empirical efforts.

The study adopts Borsboom et al.'s realist definition of validity, which invites validation research to focus on the process level of taking a test in order to explore how variation in the attribute (the construct) leads to variable outcomes (test results). It draws on qualitative data from the points of view of the test-takers, employing a think-aloud technique and stimulated recall interviews. The informants (n=18) were purposefully and randomly sampled from a group (n=121) of year 9 students in German schools.

Two extreme subsamples were formed with regard to test-takers' general L2 language ability as assessed by their scores on a C-Test. Subjects were asked to think aloud while they were solving the multiple choice-items of the listening test. Stimulated-recall interviews were used to look into comprehension problems and aspects of item difficulty as perceived by the test-takers.

The verbal report data were primarily coded with regard to categories which had originally informed the development of the test construct. Construct-relevant and -irrelevant processes were analysed with regard to their distribution across the two subsamples and their relative contribution to correct and incorrect item responses.

### **June Eyckmans *Translation assessment within a common European framework?***

In the field of translation studies, there is an emerging awareness of the need to obtain evidence for the quality of tests used to measure translation competence. To date, translation tests have been informed by practice rather than by empirical research and questions regarding the reliability of the assessment methods have remained unanswered.

According to a recently developed method (Anckaert et al. 2006; Eyckmans et al. 2009) translation performance indicators can be related to the underlying translation competence in a psychometrically controlled way. This norm-referenced method is said to bridge the gap between language testing theory and the epistemological characteristics of translation studies by selecting text segments with discriminating power through a process of pre-testing and item calibration. The dissemination of this method ties in with the European call for test standardisation in the sense that it allows a reliable and valid certification of translation competence across Europe. It also permits an exploration of text robustness so that tests can be validated for different language combinations.

In our paper we will report an experiment in which three methods for assessing translation competence are compared. The central research question centers on the reliability and the validity of these methods with reference to each other and homes in on the quintessential issue of text independent measurement of translation competence. The results of the experiment clearly indicate that collaboration between translation teachers and test developers is indispensable if we want to rise to the methodological challenge of achieving equivalent standards across languages.

### **Tony Green *Diving into the C levels. Towards reference level descriptions for English***

The Common European Framework of Reference for Languages (CEFR) has, since its publication in 2001, rapidly established itself as a valuable tool for comparing assessment results from diverse systems. However, the CEFR is designed to operate across languages and acknowledges that more detailed specifications or 'reference level descriptions' (RLDs) will be required to adequately demarcate levels for specific languages.

This paper will report on one strand of a large-scale project designed to develop RLDs for the C levels. This strand of the project attempts to identify language functions or speech acts that appear for the first time at the C levels in English language learning materials. A database of materials (addressing both language production and reception) has been assembled. Materials include internationally popular ELT course book series for high proficiency learners, national and local curricula/ syllabuses from a range of countries, English language test specifications and proficiency scales. An iterative process of analysis and synthesis of these sources indicated functional uses of language that have come to be identified over the past 30 years of educational practice with higher levels of proficiency.

A draft list of the language functions available to higher proficiency language learners has been synthesised from the database to contribute to the RLDs for the C levels of the CEFR. The current stage of the project involves the validation and scaling of the list of functions through a survey of English language educators around the world. Emerging findings are reported.

**Jenny Liantou *A closer look at strategic reading as a predictor of text difficulty***

This paper reports on an exploratory study that aimed at investigating the relationship between strategy use and test-takers' perceived level of reading comprehension difficulty with particular reference to the Greek State Certificate of English Language Proficiency exam (KPG). More specifically, the objective of the study was to examine whether specific reading strategies (such as rereading, note taking, guessing the meaning of unknown words, translating into L1, underlining or selectively reading and combining information from different parts of a text), exercise a systematic influence on test-takers' perception of module difficulty and may have an impact on their actual performance in the exam. As such, it constitutes part of ongoing doctoral research on the effect specific reader and text variables have on text comprehensibility. Data from a survey conducted on a national scale in the form of questionnaires administered to candidates sitting for the KPG exams will be presented (7.250 questionnaires administered during 2006-2008 examination periods) along with a discussion on test-takers' beliefs, cognitive processes and strategies employed when interacting with the reading texts and tasks. Multivariate analysis of variance has shown that frequent use of problem-solving reading strategies correlates significantly with perceived text difficulty, whereas support-type reading strategies (such as underlining text information or selectively reading and combining information from various parts of a text) are less often employed, regardless the perceived difficulty of a text. Finally, implications of the study for test validity and future strategy-based research are discussed and practical consequences for task design are considered.

**Thomas Eckes *Looking Beyond Rater Cognition: Operational Rater Types in Writing Performance Assessment***

Rater variability poses a major threat to the validity and fairness of writing performance assessments. Even extensive rater training sessions usually do not reduce rater variability to any significant degree. Viewed from a rater cognition perspective the perception and interpretation of scoring criteria play a key role in accounting for rater variability. Following a brief review of rater cognition studies, the presentation focuses on a recent study of rater types in writing assessment. Each rater type was characterized by a distinct pattern of perceived importance of routinely-used scoring criteria. A discussion of some of the limitations of that research leads to asking whether, and how, rater cognition is related to rater behaviour in operational scoring sessions. The core issues studied are: Can "operational rater types" (ORTs) be identified, with each type showing a distinct rater-by-criterion bias pattern, and are the "cognitive rater types" (CRTs) that emerged in the earlier research linked to ORTs? Specifically, the hypothesis is that criteria perceived as highly important will be rated more harshly than criteria perceived as less important. Many-facet Rasch measurement is employed to analyze rater bias patterns with respect to the operational use of scoring criteria. Rater bias measures are then cluster analysed to identify ORTs. At least for two of three CRTs, the hypothesis is confirmed, thus attesting to the existence of a link between rater cognition and rater behaviour. Implications for rater training and rater monitoring are discussed.

**Riikka Alanen, Ari Huhta *Combining language testing and second language acquisition research – insights from Project CEFLING***

Recent reviews of research on language learning and language assessment reveal lack of co-operation between the two fields. Combining their strengths, however, could help to address problems which are difficult to tackle from one point of view only. Yet how does such cooperation between language testers and SLA researchers take place in practice? What are the benefits for both parties and what are theoretical and methodological issues that need to be considered?

In the following, we will report on the project CEFLING (funded by Academy of Finland in 2007-09) set up to study how L2 English and L2 Finnish learners writing proficiency develops, in linguistic terms, across Common European Framework levels. In the project, roughly 3500 texts were collected from students in grades 7-9 (aged 12-16) by using a set of communicative writing tasks as data elicitation device, and rated with two scales, a CEFR can-do scale and a Finnish application of the CEFR scale.

We will begin by discussing the methodological and theoretical issues involved in the design of L2 tasks suitable for investigating both learning and assessment, including the type and rating of tasks. We will end by discussing the benefits of a language testing perspective to an SLA study, e.g., how serious attention to assessment improves the validity of the findings. We will also discuss the contribution of SLA to language testing, e.g., by improving the understanding of the constructs assessed and the more precise description of linguistic development needed in e.g. rating scales for diagnostic purposes.

### **Anne Dragemark Oscarson *Collaboration in understanding results – Self-assessment of EFL writing***

The aim of the study was to explore students' and teachers' experiences of integrating self-assessment into everyday classroom practice. It is based on the theory that metacognitive skills (i.e. self-regulation and self-monitoring) are important for the development of autonomous learning skills. Self-assessment may thus help to develop lifelong language learning skills as well as further the development of more comprehensive and thereby fairer assessment practices.

The study is part of a larger study, where 102 Swedish upper secondary students and two teachers participated during one school year. Forty-one of the students were interviewed in focus groups after a classroom assignment, where the writing process approach was coupled to self-assessment questions and non-corrective feedback from the teacher to help the students become more aware of their language skills and language levels. The two participating teachers were interviewed individually after the larger study was completed.

Students and teachers were positive to the incorporation of self-assessment practices in the EFL writing classroom and saw it as a transferable skill that supports lifelong learning, also in other subject areas.

The method used was found to be a practical way of helping students become more aware of their language skills and language levels. Both teachers and students considered student self-assessments as contributing valuable additional information to ordinary tutoring and testing.

The implications are that syllabus goals that encourage student responsibility and autonomy through for example self-assessment are viable and realistic, but students need to practice self-assessment to become adept at employing the approach.

### **Jose León, Inmaculada Escudero *Assessing summaries from Secondary School students comparing human graders to LSA in different texts***

In the present study, we tested a computer-based procedure for assessing very concise summaries (50 words long) of two types of text (narrative and expository) using Latent Semantic Analysis (LSA) in comparison with the judgments of four human experts. LSA was used to estimate semantic similarity using six different methods: four holistic (summary–text, summary–summaries, summary–expert summaries, and pregraded–ungraded summary) and two componential (summary–sentence text and summary–main sentence text). A total of 390 Spanish middle and high school students (14–16 years old) and six experts read a narrative or expository text and later summarized it. The results support the viability of developing a computerized assessment tool using human judgments and LSA, although the correlation between human judgments and LSA was higher in the narrative text than in the expository, and LSA correlated more with human content ratings than with human coherence ratings. Finally, the holistic methods were found to be more reliable than the componential methods analyzed in this study.

### **Dina Tsagari *Linking textbooks to the CEFR: a collaborative approach***

Ever since the publication of the CEFR (Council of Europe, 2001), various educational providers such as examination designers, textbook publishers, and curriculum developers make claims about the relationship between their products and the CEFR. Such claims of links have led to the production of enormous amounts of books, exams, and curricula in various educational contexts around Europe and beyond. However, there is little empirical evidence to back up claims of linkage to the CEFR.

The present research examined the claims of textbook publishers of a series of new EFL books recently introduced in the Greek State school system for Levels A1 to B1. Pre-use and in-use analysis of the textbook materials was undertaken by two groups of EFL teachers using a series of checklists examining the nature and use of texts, tasks and other textbook features and linking its contents to the targeted CEFR levels. The checklists were mainly informed by the CEFR Content Analysis Grids (Manual for

Relating Language Examinations to the CEFR, 2009:153-179). In their groups, teachers had to collaboratively identify and categorise texts and tasks and reach agreement as to the CEFR level of the materials. The results of the study revealed interesting findings about the nature of the textbooks and the ways in which writers chose and designed the textbook materials in their attempt to conceptualize the desired CEFR levels. The presentation will conclude with a discussion of the benefits of teacher collaboration in textbook analysis and the ways material writers need to approach the task of designing textbooks linked to the CEFR.

***Evelina Galaczi Collaboration and symbiosis: Using statistics and expert judgement in rater training/standardisation materials***

It is a widely accepted premise in L2 assessment that scoring validity plays a fundamental role in ensuring test fairness. Scoring validity in a speaking test is a complex issue, with rater variability playing a key role. Such rater variability (e.g., raters who are excessively harsh/lenient or inconsistent) is typically minimised through rigorous rater training and standardisation. Situated within this broader context of scoring validity, the presentation focuses on the process adopted by a large examinations provider in developing training/standardisation materials for oral examiners. The presentation will begin with a brief overview of issues relevant to scoring validity and examiner training/standardisation. It will then discuss the role of two complementary sources of information in developing rater training/standardisation materials: (a) the use of Multi-facet Rasch measurement in generating 'fair average marks', and (b) the use of expert judgement in addressing problematic marks and selecting exemplar performances. The research design for the marks collection will be described, as well as the different stages at which expert judgement is considered. The main points will be illustrated with examples from past rater Raining/standardisation materials. The presentation will end with a discussion of the value of collaboration at various levels: at the macro level through the sharing of ideas and expertise (as seen in this overview of practices adopted by one large-scale examinations provider and their relevance to those involved in the training and/or standardisation of oral examiners), and at the micro level through the 'collaboration' between multiple sources of methods and data.