Research Notes

Contents

Editorial Notes

Welcome to issue 34 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

This issue is the first of two issues that continue the theme of the ALTE 2008 conference that Cambridge Assessment hosted in April. This event focused on the social and educational impact of language assessment and involved hundreds of delegates from many countries. In this issue we include contributions from Cambridge ESOL colleagues and external contributors, all of whom describe various aspects of the educational impact of language assessment in a range of contexts. Alongside *Research Notes* issues 34 and 35 an ALTE 2008 Proceedings volume is being prepared for the Studies in Language Testing series which will be published in 2010.

In the first two articles, the impact of society on language testing is explored in two contrasting contexts. Cecilie Carlsen describes the impact of Norwegian society on the development of National Tests of English for school children. Next, Jessica Wu investigates the impact of high-stakes testing in Taiwan from the perspective of students and teachers in relation to a popular general English test.

The following pair of articles describe recent developments in Cambridge ESOL's language assessment products relating to the computer-based testing of writing and a new Teaching Knowledge Test module focusing on CLIL. Lucy Chambers explores the impact of computer-based formats on writing performance by comparing the paper-based and computer-based writing from the PET exam. Mick Ashton and Evelina Galaczi then describe the background to the new CLIL module for Cambridge ESOL's Teaching Knowledge Test, a flexible way to assess teachers' knowledge about teaching a non-language subject (e.g. history or science) through a second or foreign language such as English.

We next focus on the impact of specific exams, with case studies from Uruguay, Poland and Romania. Gerardo Valazza reports on an impact study of TKT on teachers and schools in Uruguay, considering in particular teacher learning. Next, Jo Lewkowicz and Elźbieta Zawadowksa-Kittel present a survey of teachers preparing students for the new school-leaving examination of English in Poland, focusing on perceptions of how it is affecting their teaching and their students' ability to communicate in English. Turning to Romania, Codruţa Goşa and Luminiţa Frenţiu describe a questionnaire-based case study into teacher and student attitudes towards two high-stakes English language examinations: the English tests of the Romanian school-leaving examination – the Bacalaureat – and Cambridge ESOL's Certificate in Advanced English.

In the final article Sacha DeVelle describes a study to examiner raters' use and perceptions of the revised IELTS Pronunciation scale.

We finish with short reviews of recent publications of interest.

Editorial team for Issue 34: Fiona Barker, Kirsty Sylvester and Lynda Taylor.

The role of testing in an egalitarian society

CECILIE CARLSEN NORSK SPRÅKTEST, UNIVERSITY OF BERGEN

Introduction

Test impact on society and individuals has been the subject of considerable research interest in the field of language testing during the last two decades (Alderson and Wall 1993, 1996, Bailey 1996, Shohamy 2001, Wall and Horak 2006, 2007, 2008). The social consequences of test results are regarded as a central aspect of construct validity according to Messick's definition (Messick 1989), and the focus on test impact is claimed to distinguish modern language testing in the communicative paradigm from language testing before the 1970s (Bailey 1996).

When the relationship between testing and society is explored in our field, the focus is almost exclusively on the impact *of* testing *on* society. It does however seem a reasonable assumption that there is a two-way relationship between testing and society: not only do language tests affect society; language tests are also affected *by* society. The kind of society of which the tests are a part affects test development, testing policy, the use of tests, as well as public opinion about tests. In this article the impact of society on testing will be discussed using Norway and the development of national tests for Norwegian school children as an example. I will start by describing Norway as an egalitarian society with strong socio-democratic traditions. Thereafter I will describe the Norwegian school system as a means to achieving equality and social mobility, and finally the role of testing within this system will be discussed.

Norway – an egalitarian society with strong socio-democratic traditions

Norway and other Scandinavian countries are often described as examples of well-functioning welfare states, referred to as The Scandinavian Model. This model is usually associated with:

> '[...] the State providing, financing and regulating welfare services for all citizens from cradle to grave. It is assumed to be the successful accomplishment of a strong and well organized labour movement of social democratic inclination; and it has been understood as a third or middle way between capitalism and communism.' (Abrahamson 1999)

Important social tasks, such as education and healthcare, as well as care for children and the elderly, are regarded as the state's responsibility, catered for at the public expense. There is low social inequality and poverty, a high level of employment including high female employment, gender equality, small differences in wages, and a tax policy based on the principle of condition, meaning that the more you earn, the higher the taxes on income. It has been a political aim to smooth out social differences, and to promote mobility between socio-economic classes: everybody should have the same rights and the same opportunities to succeed in life regardless of their background.

This equality has an economic as well as a cultural side: when compared with the other OECD-countries with regard to differences in salaries between higher and lower earners, Sweden, Denmark and Norway are the three countries with the smallest differences, while USA, Canada and Portugal have the greatest differences (Samlerapport–Kunnskapsdugnad for verdiskapning 2007). The policy of economic equality is strongly supported by public opinion: investigations show that despite the existing small differences in wages, 70% of people want these differences in salaries to be reduced even further (Langeland and Stene 1999).

Investigations into people's perception of the society in which they live also show that most Norwegians believe that their society is one in which there is a fair distribution of economic resources. Fifty-six percent of Norwegians describe their society as one in which most people are to be found in the middle of the social pyramid. By comparison, only 12% of the French describe their society in the same way. Indeed, almost 50% of the French describe their society as a classical pyramid, having a small elite on top, many in the middle, but most on the bottom. This is also the view of most North Americans when describing their own society (Hjelbrekke and Korsnes 2006).

In addition to the economic equality described above, there is what we could call a principle of cultural equality in Norway; a preference for the average and a dislike for the extraordinary, i.e. people who are, or think they are, better, wiser, more clever etc. than the rest. The corporate "Crew person of the month"- kind of ranking is alien to the Norwegian way of thinking. People rather tend to be modest about their achievements and to play down their deeds. This way of thinking is captured in a concept from literature, the so-called "Janteloven" (Jante Law), described by the Norwegian/Danish author Aksel Sandemose in 1933 and presented in the novel *En flygtning krydser sit spor (A Fugitive Crosses his Tracks)*. Janteloven's main tenet is summed up in the phrase: 'Don't think you're anyone special or that you're better than us', and some of its rules are: '1. Thou shalt not believe thou *art* something' and '6. Thou shalt not believe thou art greater than *we*'. The law describes a social phenomenon where people do not want to differ far from the norm. The Jante Law keeps people in their place. The phenomenon has its equivalent in the "Tall Poppy Syndrome" in the UK, Canada, Australia and New Zealand, though I believe, its influence has fallen far short of that of the Jante Law in Norway and Denmark.

The Norwegian school system

Norway has a long tradition for a "unitary school", which is a strong state-run, public, non-paying, anti-elitist school, where children from different social backgrounds meet in the same classroom. There are relatively few paid-for

private schools in Norway, partly because of rather strict laws regulating this area. Less than 2% of Norwegian school children attend private schools (as opposed to roughly 15% of French, 12% of Danish, 10% of North-American and 7% of British school children). Because of the principle of equality, there are no marks in primary school (for ages 6–13), as one of the aims of education is to differentiate between children as late as possible in their schooling.

It has also been an important aim of the Norwegian school system to wipe out any social differences, i.e. the school has been instrumental in promoting social mobility and socio-economic equality. This aim is explicitly stated in the introduction to the School Curriculum of 1997:

> 'Equality should be safeguarded between urban and rural areas, between social classes, genders, generations, between ethnic and linguistic groups and minorities, and across capability of functioning and across the range of abilities.' (Curriculum L97; own translation)

In order to fully comprehend the meaning of this quote, it is necessary to take a closer look at the concept of *equality*. The concept may be defined in different ways and given somewhat different interpretations. The Norwegian professor of sociology and minister of education from 1990–95 Gudmund Hernes (1974) distinguishes between four kinds of equality in relation to education: *Formal equality, resource equality, competence equality* and *result equality*. The first of these, *formal equality*, means that everybody should have the same chances to enter higher education, regardless of factors such as race, gender etc. There should be no differences *formally* as to who may access higher education. Yet, as long as parents have to cover the expenses, children from less-advantaged homes will not benefit from this formal equality, and inequality and social differences between classes will be reproduced. The second kind, *resource equality*, means that the socio-economic situation of the parents should not influence their children's opportunities in life. Since not all parents are able to pay for their children's education the State should give financial support by providing a student loan and different kinds of scholarships. According to this principle of equality, everybody gets the same financial support – but again, Hernes claims, inequality is the result. Resource equality gives everybody a chance to participate in the same competition, but it does not compensate for differences in children's background. The third kind is *competence equality*, which means that more public finances are used on higher levels in education than on lower levels. The more effort a pupil makes, the more support he or she obtains. Clever pupils eventually receive more as they proceed in the educational system than the less clever students who drop out earlier, start to work and pay taxes which, in turn, finance the studies of the clever students. Again, inequality is the result. The fourth kind of equality, and the one which according to Hernes is the only kind that truly promotes social mobility, is *result equality*. The school system should not only give everybody the same chances, but compensate for differences in social background, by giving more to those with less socio-economic resources. This principle is reflected by Roemer who argues that:

> 'The ideal of the equal opportunity policy is to allocate educational resources to render it so that how well a person eventually does in the acquisition of the outcome in question reflects only his effort, not his circumstances.' (Roemer 2000:23)

To achieve this goal, more financial support is needed for pupils with a less-advantaged socio-economic background, in order for them to reach the same level of competence as the more economically-advantaged child. As Hernes argues, 'Equality in results is ensured by inequality in the resources directed towards each pupil' (Hernes 1974; own translation).

The role of testing within the unitary school system

What is the role of testing within a school system which aims for result equality and social mobility? Does testing have a role to play in a society where there is a strong preference for mediocrity, and where the Jante Law makes people unwilling to stand out as excellent? Norway is of course a meritocratic society in the sense that scholarships, attractive jobs, and positions are distributed on the basis of qualifications and not on birthright. In society, then, testing has a role to play in achieving a fair distribution of privileges. But in the school system, and particularly in primary school, there has traditionally been very little testing as compared to other European countries. In primary schools, the ideal of late differentiation between pupils has been dominant with no marks being given until secondary school (age 13). As a rule, testing in primary school has been limited to two kinds: firstly, to control whether pupils have done their homework, and secondly, standardised tests which aim to identify pupils with reading or writing difficulties or other kinds of learning disabilities. The principle of result equality requires a means to detect which pupils are in the need of more resources in order to obtain the same results as their peers. The standardised tests indicate where the extra resources are needed to give everyone an equal opportunity to perform well. Since the purpose of the standardised tests has been to identify pupils with learning difficulties, the tests are constructed to discriminate between the weak pupils and the others, but not between the average and the clever pupils, or between the clever and the very clever pupils. Consequently, primary school pupils, teachers and parents are not used to a kind of testing that challenges even the clever pupils.

The best school system in the world – or not?

In Norway it has been the ambition of different political parties, Social Democrats and Conservatives alike, to have the best school system in the world. Before 2000, the general opinion in Norway was that our school system was indeed a very good one. The results of the Programme for International Student Assessment 2000 (PISA) radically changed that view: PISA 2000 compared the reading skills of 15 year-olds in 15 OECD countries. Norway performed averagely, just a little better than the OECD-mean (505 vs. 500 points), but not as well as Sweden (526 points), and far behind the PISA winner Finland (546 points). There were also positive findings, however: for instance Norwegian school children obtained high scores on social well-being at school, a finding supported by the UNICEF 2007 report:

An overview of child well-being in rich countries. Almost 40% of Norwegian school children aged 11, 13 and 15 reported that they "like school a lot". Actually, Norway is on the very top of the OECD countries when it comes to well-being at school, and interestingly, Finland, the PISA-winner, is at the bottom. Only 7% of the Finnish school children reported that they "like school a lot". These positive results received little attention, however, perhaps due to political changes which will be discussed further below.

Shift of government, shift of school policy

In 2001 a Conservative government replaced the Social Democratic government. The Conservatives wanted a shift in school policy, and they used the negative PISA results as a justification for change. The PISA results showed clearly, they claimed, that the Norwegian unitary school was not good enough, and the laws regulating the area of private schools were liberalised, making it easier to establish private schools (Friskoleloven 2003). The Conservative Government wanted competition between schools, and they needed a means of informing the public about school quality, giving parents the possibility to choose the best schools for their children, private or public. Through the introduction of a new curriculum the Conservative Government gave more freedom to teachers as to the content and methods of their teaching, but at the same time, introduced more control of learning outcomes. In other words there was less control of the input of teaching, but more control of learning output (Lieberg 2007). These two aspects, the need to inform the public about school quality on the one hand, and the need to control learning output on the other, allowed the Government to implement more testing in school, including primary school.

National tests for Norwegian school children

The Conservative government took the initiative to develop national tests for Norwegian school children in 2003. The proposal gained support in Stortinget (the Norwegian Parliament; St.prp. nr. 1 Tillegg nr. 3, 2002–2003). The tests were to fulfil two different functions (Hasselgreen, Moe, Carlsen and Helness 2004):

- Pedagogical function: the tests should 'provide pupils, teachers and administrators with the necessary information to facilitate pedagogical development.'
- Reporting function: the tests should: 'provide local and national authorities and the general public with information which can encourage dialogue and development of education standards.'

It soon became obvious that there was strong tension between the two aims. For the test developers the pedagogical aim was of paramount importance, which meant constructing tests with positive washback effect, yielding useful and detailed feedback to teachers and pupils. This concern for the pedagogical consequences conflicted with the concern for the reporting function of the tests: scores on school level were to be easily reportable on the internet, which meant that single scores were preferred to profiles and detailed feedback. In addition, in an early evaluation of the tests, high reliability estimates were given more importance than positive pedagogical consequences of the tests (Lie, Caspersen and Björnsson 2004, Carlsen, Hasselgreen and Moe 2004).

Children were to be tested at four points in primary, secondary and upper secondary school, at grades 4, 7, 10 and 11, which means that they would be 9–10 years old the first time and 16–17 the last time they were tested. National tests should be developed in four basic skills: reading in L1, writing in L1, English (reading and writing) as well as arithmetic. In the remainder of this article I will concentrate on the development of the English tests.

National tests in English for Norwegian school children

The development of the national tests in English started out with strong pedagogical intentions within the team of test constructors.[1] The national tests received a great deal of public attention, and they were felt to be high-stakes for teachers and school owners, so that the washback effect on teaching could be assumed to be strong, according to Alderson and Wall's *Washback Hypotheses* (Alderson and Wall 1993). The concern for positive washback effect on teaching and learning was considered of paramount importance to the group constructing the tests. The government initially wanted computerised tests that could be scored objectively, but out of concern for the washback effect, the test constructors insisted on testing written production as well on all levels except in the 4th grade, since the teaching of English on this level is primarily focused on oral skills. For the other grades, writing tests consisting of three different tasks were developed. The assessment was based on a rating grid reflecting models of communicative competence, and based on the CEFR levels. The rating grid was to be used as a guide for teachers when giving pupils and parents feedback about individual pupils' strengths and weaknesses. The politicians decided that it would be too expensive to pay a group of trained and experienced raters to rate the essays, so they wanted teachers themselves to rate their pupils. The test developers warned the politicians about the negative consequences of this procedure for the reliability of test scores, but this was not taken into consideration. However, the positive side-effect of this procedure was that all of the English teachers became familiarised with the CEFR and received training in assessing writing, an advantage much appreciated by the test developers. This was felt to be particularly important in primary schools where about 50% of the teachers have no formal education in teaching English whatsoever (Drew 2004, Lagerstrøm 2007). Many teachers were therefore happy to learn more about assessing writing, and felt that the CEFR-based scale was a helpful tool in assessment as well as in informing pupils and parents. Another positive side-effect was that pupils in the 7th grade started to practise writing in English, something they had done very little of prior to the introduction of the tests.

1. The author was part of the team that developed the national tests in English from their start in 2003 until March 2005.

In addition to the writing test, a computerised reading test was developed. This test was adaptive on test-level, meaning that pupils first got a series of items, and depending on their performance on these items, they were presented with a main test at a difficulty level adapted to their level of proficiency. This meant that pupils in the same class would get somewhat different tests: the strong pupils would get a chance to perform to the best of their ability, and the weaker pupils would not have to be confronted with a series of items too difficult for them. A lot of work was put into developing a large item-bank, piloting test items and standard-setting items to the CEFR levels.

Public reactions to the national tests

The introduction of national tests of writing, reading, English and arithmetic received a strong negative public reaction amongst teachers, parents and pupils, who organised action groups in order to put a stop to the tests. Pupils, particularly at upper secondary schools, boycotted the tests by staying away from school on the days the tests were administered, and their boycott gained support from parents' action groups.

It is perhaps not surprising that pupils are negative to testing in general, and about being subjected to a system of tests based on a political decision in particular. What is interesting about the negative reactions that arose is therefore not the negative response in itself, but the arguments that were used. The negative reactions were mainly due to the egalitarian ideal: there was fear that publishing the results on the Internet would lead to the establishment of more private schools for those who could afford them, and an impoverishment of the state schools. In short, the negative reactions were to a large extent based on a fear that the national tests would contribute to increased differences between the rich and the poor (Elevaksjonen Bergen 2004).

In addition, teachers feared the extra workload involved, while parents and pupils feared that the national tests would increase existing pressure on students. There were also some critical voices raising the question of a negative washback effect on teaching and learning, though the criticism showed little awareness that washback can be positive as well as negative, depending on the test itself and on teaching prior to the test.

Many primary school teachers also claimed that the tests were far too difficult, even though the piloting of test items clearly showed that this was not the case. This is probably due to the fact that the other standardised tests known to teachers were the diagnostic tests referred to above, whose main function was to identify pupils with reading and writing difficulties. Primary school teachers, parents and pupils are used to tests where the average and the clever pupils get everything right and take fright at tests that are challenging for the stronger pupils.

Current testing situation in Norway

Globalisation has led to an increase in international student assessment programmes such as PISA. The results of such studies have made it clear to Norwegian politicians that a system of quality control of learning output needs to be used on a regular basis. Even though we currently have a Social Democratic Government with a school policy quite different from the one which introduced the national testing system, the national tests are still developed and administered. Politicians acknowledge the need to assess children's basic skills of reading, writing, English and arithmetic, and realise that this cannot be done without tests which discriminate not only between the weak pupils and the others, but also between the average and the clever pupils.

Conclusion

The main concern of this article has not been to criticise or to defend the Norwegian national tests. Nor has it been my concern to defend or criticise those who oppose the introduction of the national tests. My main concern has been to demonstrate that testing in an egalitarian society like Norway is faced with particular challenges. There is a large degree of opposition towards testing and grading in a society where equality is the aim. This is something of a paradox, testing being a crucial part of a democratic society. In a society where goods, positions and privileges are distributed by qualifications and not birthright, testing is an indispensable tool.

The traditional public opposition towards testing in Norway has acted as a brake on the professionalisation of the field of language assessment and testing. It is still difficult to raise a professional debate regarding test quality, test ethics and a fair and reasonable use of test results in Norway. Rather it tends to stagnate in a discussion for or against testing altogether (Carlsen 2008). The conclusion of this article, then, must be that international testing organisations such as ALTE and EALTA and their members have an important role to play in raising the public's consciousness about testing in Europe from parents and teachers to politicians.

References and further reading

Abrahamson, P (1999) The Scandinavian model of welfare, in Bouget, D and Palier, B (Eds) *Comparing Social Welfare Systems in Nordic Countries and France*, Paris: MIRE.

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

— (1996) TOEFL preparation courses: a study of washback, *Language Testing*, 13/3, 280–297.

Bailey, K (1996) Working for washback: a review of the washback concept in language testing, *Language Testing*, 13, 257–9.

Carlsen, C (2008) Er testing skadelig? *Kronikk, Bergens Tidende*, 13.02.2008.

Carlsen, C, Hasselgreen, A and Moe, E (2004) *Kommentar fra Engelskprosjektet til rapporten "Nasjonale prøver på prøve"*. Retrieved from http://www.regjeringen.no/nb/dep/kd/pressesenter/pressemeldinger/2004/Kommentar-fra-Engelskprosjektet-til-rapporten-Nasjonale-prover-pa-prove.html?id=103221

Curriculum L97, Læreplanverket for den 10-årige grunnskolen (L97) (1997) Oslo: Kirke-, utdannings- og forskningsdepartementet.

Drew, I (2004) *Survey of English teaching in Norwegian primary schools*, Stavanger: Stavanger University College.

Friskoleloven (2003), retrieved from http://www.regjeringen.no/nb/dep/kd/dok/rundskriv/2003/Informasjon-om-den-nye-friskoleloven.html?id=109340

Hasselgreen, A, Moe, E, Carlsen, C and Helness, H (2004) CATwalk to quality: The painstaking process of developing CEF-based computer-adaptive reading tests in Norway, paper presented at *EALTA Conference*, Kranjska Gora, Slovenia, May 14–16, 2004.

Hernes, G (1974) Om ulikhetens reproduksjon: hvilken rolle spiller skolen?, in Mortensen, M (Ed.), *I forskningens lys*, Oslo: NAVF, 147–170.

Hjelbrekke, J and Korsnes, O (2006) *Sosial mobilitet*, Oslo: Det Norske Samlaget.

Lagerstrøm, B O (2007) *Kompetanse i grunnskolen. Hovedresultater 2005/2006*, retrieved from http://www.ssb.no/emner/04/02/20/rapp_200721/

Langeland, O and Stene, R J (1999) *Holdninger til arbeid, lønn og fagbevegelse Resultater fra en spørreundersøkelse*, Rapportserien i prosjektet Det 21. århundrets velferdssamfunn, Oslo: FAFO.

Lie, S, Caspersen, M and Björnsson, J (2004) *Nasjonale prøver på prøve*, retrieved from http://www.utdanningsdirektoratet.no/templates/udir/TM_Artikkel.aspx?id=1394

Lie, S, Kjærnsli, M, Roe, A and Turmo, A (2001) Nasjonal hovedrapport PISA 2000: Godt rustet for framtida? Norske 15-åringers kompetanse i lesing og realfag i et internasjonalt perspektiv, in *Acta Didactica No. 4*, Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.

Lieberg, S (2007) Den nye læreplantenkningen – hvilke utfordringer og muligheter gir den i opplæring av voksne, paper presented at VOX-conference: *Faglige dypdykk – læreplanen fra flere sider VOX-konferansen om norskopplæring for voksne innvandrere*, Oslo, 14–15 May 2007.

Messick, S (1989) Validity, in Linn, R (Ed.) *Educational Measurement*, New York: American Council on Education.

Roemer, J (2000) Equality of opportunity, in Arrow, K, Bowles, S and Durlauf, S (Eds.) *Meritocracy and Economic Inequality*, Princeton: Princeton University Press.

Samlerapport – Kunnskapsdugnad for verdiskapning (2007) retrieved from http://www.kunnskapsdugnad.no/rapporter.html

Sandemose, A (1933) *En flygtning krydser sit spor*, Oslo: Tiden Norsk Forlag.

Shohamy, E (2001) *The Power of Tests. A Critical Perspective on the Uses of Language Tests*, London: Longman.

Stortingsproposisjon nr. 1 Tillegg nr. 3 (2002–2003) *Nasjonalt system for kvalitetsvurdering i grunnopplæringen*, retrieved from http://www.regjeringen.no/nb/dep/kd/dok/NOUer/2003/NOU-2003–16/19/5.html?id=370806

UNICEF (2007) *Child poverty in perspective: An overview of child well-being in rich countries*, retrieved from http://www.unicef-irc.org/publications/pdf/rc7_eng.pdf

Wall, D and Horak, T (2006) *The TOEFL Impact Study: Phase 1. The Baseline Study*, TOEFL Monograph 34, Princeton, NJ: Educational Testing Service.

— (2007) Using Baseline Studies in the Investigation of Test Impact, *Assessment in Education,* 14/1, 99–116.

— (2008) The TOEFL *Impact Study: Phase 2. Coping with Change*, TOEFL iBT Research Series, No. 05, Princeton, NJ: Educational Testing Service.

Views of Taiwanese students and teachers on English language testing

JESSICA WU THE LANGUAGE TRAINING AND TESTING CENTRE, TAIWAN

Introduction

With the development of Taiwan's economy and the shift in Taiwan to a more international outlook, there has been a strong identification in recent years of the need for residents to acquire competency in English. This interest has been supported by government policies concerning the use of English language testing and assessment.

In 2005, the Ministry of Education (MOE) adopted the CEFR, Common European Framework of Reference for Languages: learning, teaching, assessment (Council of Europe 2001), as its source for the establishment of target levels of English ability for EFL learners in Taiwan. Currently in Taiwan, a score on an external English test is influential in enabling individuals to graduate from educational institutions or obtain job promotions.

Governmental support and developing social trends have resulted in an enormous increase in the population of English language test-takers in Taiwan. In 2007, the number of test-takers sitting for the four most popular English tests – GEPT, IELTS, TOEFL, and TOEIC – reached a record high of 600,000. Among them, the GEPT was the most widely used, accounting for over 80% of the total number of test-takers (approximately 500,000).

Context of the study

The General English Proficiency Test (GEPT) is a five-level criterion-referenced EFL testing system implemented in Taiwan to assess the general English proficiency of EFL learners. In 1999, the MOE lent its support to the Language Training and Testing Centre (LTTC) in its development of the GEPT. The aim of the GEPT is to promote the concept of life-long learning and to encourage use of the communicative approach in English teaching and learning. The test was created in response to comments from educators and employers about the general lack of ability to communicate in English among Taiwanese English learners, partly due to the previous 'old-fashioned' approach to English teaching, which over-emphasized the importance of grammatical accuracy.

Each level of the GEPT consists of four components:

listening, reading, writing, and speaking. The GEPT is being used by various government institutions and schools for entry, classroom achievement, and graduation requirements. So far, about 3.2 million EFL learners in Taiwan have taken the GEPT since its first administration in 2000.

A number of studies related to the GEPT have been conducted by the LTTC on: parallel-form reliability (Weir and Wu 2002); the concurrent validity of the GEPT Intermediate and High-Intermediate tests (LTTC 2003); mapping the GEPT to the Common English Yardstick for English Education in Taiwan (LTTC 2005); test impact (Wu and Chin 2006, Wu 2007); test form and individual task comparability (Weir and Wu 2006); and relating the GEPT reading comprehension tests to the CEFR (Wu and Wu 2007).

Tests and assessments have long been used in education as efficient tools of reform because they can be externally mandated. A minimum standard of English proficiency as a graduation requirement has largely been implemented in Taiwan's tertiary education system, where the minimum standard of English proficiency can be represented by different indicators, which include achievement in an external, formal test of English, such as the GEPT, achievement in an internal test held on a campus, and even achievement in classroom assessment during a language course.[1]

The situation described above clearly demonstrates that English language testing and assessment, including both exams given externally and formally, and assessment in the classroom, play a very important role in Taiwan's EFL education, particularly at the tertiary level. In addition, since students and teachers are immediate, important stakeholders in testing and assessment, it is desirable to give them a voice when we attempt to understand issues related to current English language testing and assessment practices in Taiwan. As part of a project on the impact of the GEPT, the present study aimed to survey students and teachers in order to determine their general attitudes towards and views on English language testing and assessment in both formal and classroom situations. Specifically, by means of a questionnaire, the study intended to answer the following questions:

1. How do students and teachers view external exams?
2. How do students and teachers view assessment practices in classrooms?
3. What are the similarities and differences between students' and teachers' views on external exams and assessment in classrooms?

Method

A questionnaire survey was used to collect opinions on current English language testing and assessment practices in Taiwan from a number of Taiwanese students and teachers, without drawing comparisons between different schools, regions, or educational levels. In light of its exploratory nature, the present study employed convenience sampling rather than stratified sampling. Thus, the findings of the survey results should be interpreted with caution.

1. For the sake of consistency in the terminology used in the present article, an external or a formal test/exam is described as an 'external exam'.

A total of 229 students took part in the survey, including 166 students from three high schools (32 in northern Taiwan; 92 in central Taiwan; 42 in southern Taiwan) and 63 students from one university in southern Taiwan. A total of 89 teachers took part in the survey, including 26 high school teachers (16 in northern Taiwan; 10 in central Taiwan) and 63 college/university teachers (46 in northern Taiwan and 17 in southern Taiwan).

Two questionnaires were employed in this study to collect reflections from students and teachers, respectively. The questionnaires were adopted from the ENLTA Project (European Network for Language Testing and Assessment; Erickson and Gustafsson 2005). The ENLTA questionnaire was also intended to collect reports from both students and teachers on language assessment in the European context. To suit the Taiwanese context, the questionnaires were translated into Chinese. The translated questionnaires were piloted among a small sample to check whether the questions were clearly worded and understood. Adjustments were made on the basis of the pilot results.

The student questionnaire contained a combination of open-ended questions and Likert scale statements. First, the students were asked to provide answers to the following open-ended questions: What is a good/bad English language test/assessment? Why? The questionnaire also included 18 statements, accompanied by 5-point Likert scales, that enquired about their attitudes towards English assessment in general and their views on specific areas of assessment. The students were asked to indicate to what extent they agreed with what was stated (5= Yes, absolutely; 1= No, absolutely not) about external exams and classroom assessment. The statements on both topics were identical in order to make some comparison between the answers possible.

The teacher questionnaire was similar to the student version. To facilitate some comparison between students' and teachers' views, the scale statements about areas of assessment and the open-ended questions on assessment were identical to those in the student questionnaire. However, there was an additional open-ended question for teachers: How have external exams (e.g. Entrance Exams, GEPT, TOEFL, and IELTS) affected your teaching and assessment practices in the classroom?

Major findings

Owing to the willingness of all the students and teachers who participated in the survey to share their knowledge, experiences and views on English language tests and assessment, some major findings were obtained and their implications are discussed in the following sections.

More positive attitude of students towards classroom assessment

Five statements were used to elicit students' affective attitudes towards testing and assessment. In this part, a distinction was introduced between 'external exams' and 'classroom assessment'. The degree of agreement with the statements is shown in Table 1.

Table 1: Students' degree of agreement with statements

Statement	Degree of agreement	
	External exams	Classroom assessment
There are opportunities for me to show what I know.	2.93	3.31
I learn a lot.	3.23	3.47
I often feel nervous.	3.63	3.07
Important things are assessed.	3.25	3.34
I often do well.	2.71	3.03

The students had more positive attitudes toward classroom assessment, but on the whole, the students' views did not distinguish very clearly between external exams and classroom assessment. Not surprisingly, an exception was observed: the mean value for the statement 'I often feel nervous' was markedly higher with regard to external testing statements than classroom situations.

It is not surprising to see more favourable attitudes of students towards classroom assessment, given the fact that tests and assessments are frequently administered in classrooms and are less formal in nature, whereas external exams are high-stake tests given in formal situations, and produce results that are likely to influence students' futures. Having said that, with a better understanding of how students may feel about external exams based on their responses in the survey, we should constantly remind ourselves of the importance of listening to students' voices about their needs in learning and assessment. In addition, to assess students' English proficiency in a fair manner, we should also look for alternative methods of assessment that can be used to determine students' learning progress and achievement.

Influences of external exams on teaching and assessment practices

Over 90% of the teachers reported that external exams influenced their teaching and assessment practices. The following comments illustrate teachers' views of this influence:

> The external exams have made me change my way of assessing my students.
>
> I'm teaching in a university, so my teaching is not as exam-driven as teaching is in high schools. However, given the minimum standard of English proficiency required of university students for graduation nowadays, I need to better understand the external exams and what they test.
>
> To help our students perform better on external exams, particularly to improve their reading comprehension ability, we recommend books for them to read. We also include listening comprehension, translation, writing, and speaking in our assessment practices in our school.
>
> External exams do influence my teaching and assessment practices. For example, I require my students to keep diaries, see English films, and guide them to talk about the films with their classmates. By doing these things, I hope to improve my students' English proficiency through a wide range of learning activities.
>
> Teaching is inevitably influenced by external exams because we do hope to see our students get better score results.
>
> Our teaching is exam-driven. The college entrance exams don't assess listening and speaking, so little attention is given to listening and speaking activities in high schools. However, since GEPT scores have been recognized as additional proof of English proficiency in the university/college application system, listening and speaking have been added to the teaching and assessment practices in my school. We often feel pressure from students who want us to provide listening and speaking exercises.
>
> That teaching and learning is exam-driven is not all bad. If an external exam can have positive impacts on teaching and assessment practices, even on education and society, then why not?

To sum up, from the teachers' reports, it is obvious that teaching and assessment practices have been influenced by external exams, namely, the college entrance exam and the GEPT. It is good to find that most of the influence is positive; notably, one of the teachers reported that he had actually learned some new ideas from the external exams, and he subsequently used them in his classroom. It is also worth noting that a few teachers reported that the full coverage of macro-skills in the external exams such as the GEPT has encouraged them to include listening and speaking in their teaching and assessment practices. Although a few teachers worried that teaching and learning were driven by external exams, they also admitted that good external exams could have positive washback effects on teaching and learning.

Similarities and differences of opinions in relation to target areas in assessment

In the survey, the students and teachers were asked to report how often they encountered various target areas in both formal testing situations and classroom situations. Comparison of their responses reveals a considerable degree of uniformity among their views.

Firstly, concerning target-assessed areas in formal testing situations, both groups reported that *reading comprehension, writing texts,* and *words/phrases* were the most frequently assessed areas, and that *pronunciation* and *speaking* were the least frequently assessed areas. Both groups strongly agreed that the areas related to speaking, i.e., *pronunciation, speaking on one's own,* and *talking with others*, were tested less often in both testing situations. This seems to contradict the finding that external exams have influenced teaching and assessment practices since speaking is covered in external exams such as the GEPT. A possible explanation for this contradiction is that a mandatory exam (e.g. the college entrance exam) that lacks a speaking component will naturally have a more powerful effect on teaching and assessment practices than will other types of external exams (e.g. the GEPT), which assess a wider range of macro-skills but are not mandatory. However, one difference was found: the students felt that *grammar* was also frequently assessed in external exams; whereas the teachers marked *others* as another commonly assessed area. This variation in responses may have been due to the fact that the groups interpreted some target areas listed in the questionnaire differently.

Secondly, there was also a good degree of consensus between the two groups on the target areas assessed in classroom situations. Both groups reported that *reading comprehension, words/phrases,* and *grammar* were the

areas most frequently assessed in classrooms. However, there was one difference between the groups: the students also identified *spelling* and *translation* as areas commonly assessed in classrooms; whereas the teachers identified *others* and *writing texts* as areas more commonly assessed in classrooms. Again, this may have been caused by differences in how the groups understood and interpreted the target areas in the questionnaire.

Strong agreement between students' and teachers' perceptions of good and bad tests/assessments

The students were highly responsive in giving comments on what was described as a good or bad test/assessment. A total of 184 students (80%) provided comments in writing. Like the students, the teachers were very willing to give their opinions on what was described as a good or bad test/assessment. Among the 89 teachers, 65 of them (73%) provided comments in writing. There was strong agreement between the students' and teachers' comments on what was described as a good test/assessment and a bad test/assessment. Such strong agreement between the groups may have been the result of teacher/student influence and students' own experiences with tests and assessments. According to both groups, five features which determine good or bad test/assessment are as follows:

- Macro skills to be assessed: A good test/assessment measures all four macro skills (listening, reading, speaking, and writing)
- Discriminating power: A good test/assessment successfully discriminates a stronger learner from a weaker learner
- Influence on teaching and learning: A good test/assessment affects teaching and learning positively
- Test conditions under which the test/assessment is performed: A good test/assessment provides appropriate conditions under which tasks are performed
- Contents of a good test/assessment include culture-related elements (according to teachers only).

Conclusion

In brief, the findings of this study have helped us to gain a better understanding of what stakeholders, in this case students and teachers, think about current English language testing and assessment (in the contexts of both external exams and classroom assessment) in Taiwan and their effects on teaching and assessment practices. Despite the limitations of the present study, this survey has demonstrated that students' and teachers' reflections deserve to be heard by other groups of stakeholders, including EFL educators, education policy makers, exam developers, school administrators, etc., in order to improve the practices and results of English language testing and assessment in Taiwan.

References and further reading

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

Erickson, G and Gustafsson, J (2005) *Some European Students' and Teachers' views on Language Testing and Assessment*, ENLTA Activity 4 Report, available at http://www.ealta.eu.org/resources.htm

LTTC (2003) *Concurrent Validity Studies of the GEPT Intermediate Level, GEPT High-Intermediate Level, CBT TOEFL, CET-6, and the English Test of Taiwan's College Entrance Examination*, LTTC Research Report, Taipei: Language Training and Testing Centre.

— (2005) *Mapping the GEPT to the Common English Yardstick for English Education in Taiwan (CEY)*, LTTC Research Report, Taipei: Language Training and Testing Centre.

Weir, C J and Wu, J (2002) Parallel-Form Reliability – A Case Study of the GEPT Spoken Performance Test, *Proceedings of the Fifth International Conference on English Language Testing In Asia*, Tokyo: The Society for Testing English Proficiency, 29–56.

— (2006) Establishing test form and individual task comparability: a case study of a semi-direct speaking test, *Language Testing*, 23/2, 167–197.

Wu, J (2002) Investigation of test-takers' views on difficulty at task level: A case study of GEPT-Intermediate spoken performance, *English Teaching & Learning* 26/4, 107–124.

—(2007) English language assessment in Taiwan: Where do we go from here? *Proceedings of 2007 International Conference and Workshop on TEFL & Applied Linguistics*, Taipei: Crane Publishing Co., 574–586.

Wu, J and Wu, Y F (2007) Relating the GEPT reading comprehension tests to the CEFR, paper presented at the *4th European Association for Language Testing and Assessment Conference (EALTA)*, Sitges, Spain, 15–17 June, 2007.

Wu, R and Chin, J (2006) An impact study of the intermediate level GEPT, paper presented at the *9th International Conference on English Testing in Asia*, Taipei, 23–24 November 2006.

Computer-based and paper-based writing assessment: a comparative text analysis

LUCY CHAMBERS RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

Introduction

In recent years there has been an increase in the use of computer technology within assessment. This has included converting traditional paper-based (PB) tests into electronic formats. Computer-based (CB) testing has a number of advantages, but there have been questions concerning score equivalence and whether the use of a different administration mode affects the test's validity. Thus it is important that issues of comparability continue to be explored.

This article summarises an exploratory research study comparing computer-based and paper-based writing assessment.[1] Unlike the majority of current research, which focuses on score comparability, this study focuses on the comparability of textual and linguistic features. The study considered whether written output produced via the two modes of administration is comparable in terms of:

- lexical resources
- organisation
- surface features
- frequency of typographical or spelling errors.

The research sought to provide insights that could inform rater training.

Technology and its effect on writing

Weigle (2002) states that technology is changing the way we think about writing and how we do it. She claims that the pervasiveness of email, chat rooms and online discussions is causing the traditional distinction between speech and writing to become blurred: 'In these contexts, writing takes on may of the aspects traditionally associated with speaking, such as shorter, less complex and even incomplete sentences, lack of attention to accuracy, and less formal language' (Weigle 2002:231). Technological changes have given rise to an increase in computer-mediated communication (CMC) in which English is a dominant language. This has meant that the Internet has become the primary medium of English-language writing for many second-language speakers around the world (Warschauer 2003:162). In addition the use of tools such as word-processors has had an impact on writing processes and thus may affect writing quality – for example, by making textual revisions easier to complete – and by the use of spell- and grammar checkers.

Comparative analyses of writing produced via different assessment modes such as in the present study can highlight whether computer technology is also affecting the nature of writing in an assessment context. Differences may be detected that reflect the influence of CMC or of different writing processes; these may have implications for rating writing assessments.

Comparability studies

Russell and Tao (2004a:3) describe two variables that may influence students' writing scores and thus comparability: the mode of composition for students and the mode of presentation for raters.

There have been a number of studies looking at mode of composition effects, particularly at direct writing score comparability and factors affecting it. These have focused on areas such as computer experience, computer anxiety, attitude towards computers, socio-economic factors and gender. In addition there have been studies looking at comparability of essay quality across administration modes, using holistic rating components such as organisation, mechanics and accuracy. The findings from these studies have been mixed and no consensus has emerged. Few studies have explored these components at a detailed text level, in order to reveal where any differences may lie.

Lee (2004) argues that findings are dependent on research contexts and specific details of the procedures. Issues that may influence findings are whether it was timed or untimed, access to grammar and spell checkers, participants having to write in an unfamiliar composition mode (be it CB or PB) and differences in motivation and attitude to composition mode. In addition, many of the studies have used raters with little professional scoring experience, such as graduate students or teachers. It must be remembered that both the rapid changes in exposure to technology and changes in technology itself mean that results from earlier studies may not be applicable today.

Russell and Plati (2002) call for candidates to be given the choice of mode in which to complete extended written assessments. They ask: 'Do we want to know how well students write on paper, how well they write on computer, or how well they write in the mode of their choice?' (cited in Horkay *et al* 2006).

Russell and Tao (2004a) state that decreasing the effect of mode of composition, by giving students the choice of mode in which to compose their essay, increases the effect of mode of presentation. Raters may be influenced, both positively and negatively, by the appearance of essays in handwritten versus typed text. Studies by Powers, Fowles, Farnum and Ramsey (1994) and Russell and Tao (2004a) found that essays presented in handwritten form receive higher scores than the same essays presented as computer text. The authors suggested a number of hypotheses to account for this: increased visibility of errors in the computer text; higher expectations for computer texts; handwriting enabling the reader to feel closer to the writer; handwritten responses being given the benefit of the doubt when hard to read; and handwritten responses appearing longer and the result of greater effort.

Examiner training and standardisation can help ensure that any presentation effects are minimised as far as possible. In a small follow-up study Russell and Tao (2004b) were able to demonstrate that the presentation effect could be eliminated with supplementary training. In order to train raters in the presentation effect and how to avoid it, we need to look at differences in appearance between the texts and how they arise.

Methodology

Authentic test data was obtained in order that the results could be directly applied to an assessment context; thus a PET part three writing task common to both a live PB and a live CB examination session was chosen. This task is a longer piece of continuous writing (100 words) in response to an extract from an informal letter written by a friend. This extract provides the topic candidates must write about, with a couple of questions included to focus their ideas. The task chosen was on the topic of sport.

The study focused on whether the written output from an assessment task was comparable in terms of linguistic and text features when produced via a paper- or computer-

1. This study was completed in partial fulfilment of a Masters degree in Applied Linguistics at Anglia Ruskin University (UK).

based administration mode. The following features were studied:

- Lexical resources: text length, standardised type-token ratio, lexical sophistication, comparison of candidate output and PET wordlists
- Organisation in terms of sentences and paragraphs
- Surface features: punctuation, capitalisation and use of text/email conventions
- Frequency of typographical/spelling errors that do and do not impede comprehension.

The methodology was finalised after a pilot study (not described here). It is important to note that the CB participants had no access to word-processing tools such as spell- or grammar checkers. Analysis was conducted using Wordsmith Tools (Scott 1998), Range (Nation and Heatley 1996) and SPSS. The linguistic features studied were based on sub-sections of the markscheme although the range and accuracy of grammar were not studied here.

Scripts were collected from the two exam sessions and candidates selected so that they were matched on exam centre and language proficiency. This meant that sampled candidates would have had the choice of administration mode thus selecting the mode most suitable to them, a condition suggested by Russell and Plati (2002:21). Thus any possible impact of computer familiarity or anxiety on the test output is likely to be reduced. Candidates from four examination centres were selected: one Columbian, two Italian, and one Swiss centre. The total number in the PB group was 86 and the total number in the CB group was 82.

Results

We present the results in four sections below, relating to lexical resources, sentences and paragraphs, surface features and spelling errors.

Lexical resources

Analysis of the two groups using Wordsmith Tools revealed that PB texts were on average 4 words longer; this group also had a greater standard deviation (19.4), suggesting greater variability between the candidates (see Table 1). A t-test revealed no significant difference in the mean number of words written between the two groups (t=-1.294, df=154, p=0.197). Thus there is insufficient evidence to suggest a difference in text length between the administration modes.

Table 1: Lexical output and variation by administration mode

Administration mode	PB		CB	
	Mean	**SD**	**Mean**	**SD**
Tokens	108	19.4	104	15.4
Types	69	11.9	69	8.7
Standardised Type/Token Ratio	77	5.4	79	5.3
Average Word Length	4	0.2	4	0.3

Many of the studies conducted on direct writing show that participants writing in CB mode generally write more (Goldberg *et al* 2000, Li 2006, MacCann *et al* 2002). These are contrary to the findings from this study. It must be remembered that this task does have a word limit: candidates are asked to write 100 words. This and the fact that it is a timed situation, will impact on the numbers of words written. In addition, the CB candidates have a word count available to them so they can be more accurate in achieving this target; which may explain the lower standard deviation for CB texts.

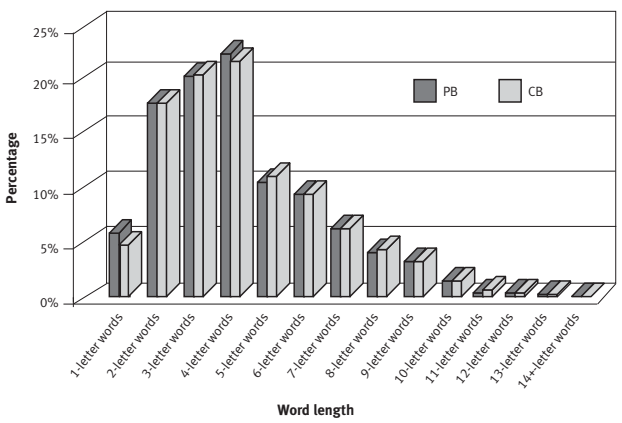
Interestingly, the mean number of types was the same for both groups (69), although there was greater variation in the PB scripts (SD 11.9) (see Table 1). This would suggest that candidates were using a similar range of vocabulary. However the PB candidates used more words; this would imply that in the PB test candidates were recycling some words rather than using the 'extra words' to add more breadth to the composition. The standardised type-token ratio confirms this; the mean ratio for the CB group is higher (79 compared with 77). A t-test revealed a significant difference in the mean standardised type-token ratio between the two groups (t=2.4, df=165, p=0.019). Thus there is evidence to suggest that the standardised type-token ratio is different across the two administration modes. However, this ratio does not provide information on the type of words used, only the relative variety of them; the analysis of lexical sophistication will shed more light on this.

Evidence from corpus studies conducted by Yates (1996) found that type-token ratios were higher for writing than for CMC. This would suggest that the standardised type-token ratio might be higher for the PB texts. This was not the case in this study so it would appear that the effects of CMC on lexical variation seen elsewhere were not evident in the CB scripts in this study. However, revisions are easier to make on computer so it could be hypothesised that in the CB mode text could have been altered to avoid use of token repetition, hence achieving a higher standardised type-token ratio.

The mean average word length is similar for both groups, suggesting a similarity in the type of vocabulary produced via the two administration modes (see Table 1). However, the standard deviation for CB scripts is higher. A t-test reveals the difference to be significant (t=2.1, df=166, p=0.039). Thus, there is evidence to suggest that average word lengths between the administrations are different. This can be explored further by looking at the distribution of different word lengths in texts.

Analysis of the mean proportion of words of different lengths appears to follow a broadly similar pattern for both PB and CB tests (see Figure 1). The PB group appears to use a greater proportion of shorter words (4 or less) than the CB group: 66% as opposed to 64%. In particular, a higher proportion of 1- and 4-letter words are used. This could account for the significant difference found in average word lengths. It would appear that there does appear to be a difference in proportion of shorter/longer words used between the modes. Examination of the types of words used should be able to reveal more about administration mode differences.

Figure 1: Proportion of word lengths used by administration mode



It is also useful to examine the lexical sophistication of the texts and to see which words are used in common and which are unique to each mode. Results of analysis using the Range programme can be seen in Table 2. This shows that the percentage of tokens in the first one thousand word list is higher for the PB group, and similar for the second and third word lists. This is not unexpected; it was established above that the PB group had a higher proportion of words shorter in length. Word length is known to roughly correlate with word frequency (Alderson 2000:72), the most common words tending to be shorter.

Table 2: The proportion of high- and low-frequency word by administration mode

Word List	Tokens %		Types %	
	PB	CB	PB	CB
One (1st 1000)	81.36	79.71	49.48	49.05
Two (2nd 1000)	8.15	8.29	8.09	8.53
Three (AWL)	0.82	0.92	2.52	2.73
Not in the lists	9.67	11.08	39.91	39.69

The percentage of types was similar across modes for all three word lists. This suggests that although there was a greater proportion of tokens in the first list for PB scripts, more of these were used with greater frequency. This appears to support the differences in the mean standardised type–token ratios found above. Perhaps the difference can be explained by the PB scripts having a higher incidence of 'recycling' of higher frequency words.

A higher proportion of tokens appeared in the 'not in the lists' category for CB texts, whereas the proportion of types was similar. The words 'not in the lists' contained a mix of proper nouns (countries and names), quite specific words related to the topic (e.g. *soccer, playoffs* and *competitive*), words that are incorrectly spelt and non-English words. The words that fell into the last two categories were classed as errors. 69% of the tokens in the CB category and 67% of tokens in the PB category of this list were not classed as errors. The majority of these in both lists were types of sport or people's names (either sports personalities or the names of the correspondents). More detail on errors can be found below.

Consistency analysis of the types used in both modes revealed that 599 types were used in common: this is approximately half the types for each mode. The types not used in common were mostly names or errors; the maximum frequency of one of these types was 9. This shows that there were not any types used extensively in one mode that were not used in the other. When looking at the differences in type frequency between modes, only 13 words had a frequency difference of more than 20, interestingly the word with the greatest difference was the pronoun 'I'.

If the fifty most frequently used types in each administration mode are compared, only 11 words differ between the modes (see Table 3, words unique to mode

Table 3: Comparison of the 50 most frequently used words in each mode

PB	Frequency	CB	Frequency
THE	401	THE	415
I	311	IN	321
IN	271	IS	247
IS	253	I	229
YOU	225	AND	229
AND	192	YOU	220
SPORT	175	SPORT	193
A	160	TO	181
TO	156	A	171
ARE	154	ARE	155
OF	145	VERY	125
MY	124	OF	124
SPORTS	121	MY	122
THAT	119	SPORTS	114
POPULAR	107	POPULAR	104
VERY	99	COUNTRY	101
BUT	97	BUT	96
COUNTRY	96	FOR	92
YOUR	96	YOUR	91
PLAY	95	FOOTBALL	90
FOOTBALL	94	IT	90
LIKE	92	PEOPLE	90
IMPORTANT	91	IMPORTANT	88
HERE	89	LIKE	80
FOR	87	WE	76
IT	84	HERE	68
PEOPLE	82	THAT	67
WE	77	THERE	67
MOST	73	LETTER	66
HAVE	67	HAVE	64
ME	59	MOST	73
BECAUSE	59	SOCCER	51
SO	59	ME	59
LETTER	57	PLAY	95
ALSO	57	CAN	50
THINK	53	DEAR	49
SOCCER	51	ALSO	57
CAN	50	ABOUT	45
DEAR	49	THIS	47
DO	48	OR	41
LOVE	48	BECAUSE	59
THIS	47	SO	59
ABOUT	45	DO	48
SEE	45	**GO**	38
ALL	43	**MANY**	38
THERE	41	**TENNIS**	34
OR	41	TOO	40
I'M	41	**GOOD**	37
TOO	40	**NOT**	25
WITH	39	**TEAM**	28

highlighted). Although the order of types is slightly different between the groups it is striking just how similar they are.

Wordsmith consistency analysis enabled comparison of the types and tokens used in the texts to the PET wordlist. Results showed that 86% of the PB tokens and 84% of the CB tokens featured in the PET wordlist. In terms of types 38% of PB types and 42% of CB types featured in the PET wordlist. Results suggest that the CB candidates used a broader range of types from the PET list but did not use them with as great a frequency as the PB candidates. This supports the evidence from the standardised type-token ratio data again implying that CB candidates are using a broader range of vocabulary.

This shows a slightly different pattern to the results of the Range analysis. The PET wordlist, whilst being a general B1 level wordlist, is also specific to PET and is updated using evidence from PET performances and corpora.[2] It contains targeted vocabulary expected of B1 users and as such contains vocabulary that will not feature in wordlists based purely on frequency, such as the General Service Lists found in the Range programme. An example relevant to the task used would be that the PET wordlist contains a variety of sports, e.g. *tennis* and *basketball*, that would not be present in the frequency-based wordlists.

Sentences and paragraphs

From Table 4 it can be seen that the mean number of sentences for the PB candidates is one sentence more than for CB candidates. There is however more variability (SD=3.5) in the PB scripts. A t-test revealed a significant difference in the mean number of sentences between the two groups (t=2.397, df=166, p=0.018). Thus there is statistical evidence to suggest that the number of sentences differs in text produced via the two modes.

Table 4: Sentence and paragraph statistics

	PB		CB	
	Mean	**SD**	**Mean**	**SD**
Sentences	8	3.5	7	2.9
Paragraphs	5	2.6	2	1.4
Words per sentence	16	15.1	16	9.6

Analysis of the data revealed that one CB candidate wrote zero sentences; in fact three paragraphs were written but without full stops. One of this candidate's paragraphs was concluded with a comma: whether this can be considered a typographical error (comma and full stop keys are adjacent on the keyboard) or lack of knowledge about punctuation we can not be sure. This however does show a limitation in using frequency statistics such as these on lower proficiency candidates in that differences may be highlighted that are due to erratic punctuation rather than differences in the actual text features that are being measured. However, punctuation is part of the marking criteria, and as such, its impact should be considered.

Punctuation also had an impact on words per sentence. The mean number of words per sentence is the same for CB and PB scripts but the CB scripts show less variation (see Table 4). However exploration of the data revealed that a number of candidates wrote very long sentences, the most extreme was a PB candidate who wrote only one long sentence: a series of clauses separated by commas. One of the CB candidates was classified as writing no sentences; the candidate did not use any concluding punctuation. It could be hypothesised that proficiency may have had an effect on punctuation and that with correct use of punctuation the statistic would have yielded different results. However, as stated earlier, punctuation is part of the marking criteria and so these findings are interesting and could provide useful feedback to teachers. It would be interesting to see both whether a sentence length effect occurred and if any differences in accuracy of punctuation existed for higher proficiency candidates.

When looking at the mean number of paragraphs between the modes (Table 4), there were on average 3 paragraphs more in the PB mode. The standard deviation is also higher suggesting more variability in this group. Examination of the data revealed that the numbers of paragraphs in the text were not normally distributed so a non-parametric Mann-Whitney test of independent samples was used to test for differences between the groups. A significant difference was found between the groups (z=-6.681, p=0.00). Thus there is evidence that mode of administration does have an effect on the number of paragraphs written.

Closer examination of the PB data revealed that some candidates wrote each sentence in a separate paragraph. Whether this was intentional paragraphing or just poor layout on the part of the candidates is unknown. From the CB group, 38% of the candidates wrote only one paragraph as opposed to just 5% in the PB group. This shows an important difference in the writing style of the two groups, and is perhaps attributable to the influence of email, where shorter emails are frequently written with the openings (if present) written in the body of the text, not in a separate paragraph. Differences in letter opening and closing between the administration modes were examined but will not be described here.

There is evidence that organisation of the texts does appear to be different between the two administration modes. This may have an impact on marking in that the presentation of the text will be different across the two modes. A single block of text compared with a text split into paragraphs is likely to affect a rater's perception of how organised the text is. Whether organisation of a text in terms of paragraphs is something that should be assessed in a test of language proficiency is a matter for discussion. It would depend on how the construct of writing was defined: i.e. whether anything other than the language used should be taken into consideration. It is important that any difference in paragraph use between modes of administration should be made explicit in both rater guidelines and training so that raters are aware of it and know how to address it.

Surface features

Errors in capitalisation could be found in scripts from both modes: 32 CB candidates and 24 PB candidates produced

2. This is the list available to item writers; a public version is available at www.CambridgeESOL.org/teach/pet

errors. There were mainly three kinds of error: omission of initial capital letters for names and countries, omission of a capital at the start of a new sentence and capitalisation at the start of words that did not require it. There were, however, 12 PB and 2 CB candidates who wrote entirely in capitals thus their appropriacy of capitalisation could not be monitored. The most striking difference in capitalisation between the modes was the use of lower case 'i'. Fourteen CB candidates and only 2 PB candidates used this feature incorrectly, suggesting that for CB candidates this is something that could be addressed. It is possible that it is caused by over-reliance of auto-correct features whilst word-processing or that writing an upper case 'I' by hand is so automatic that candidates are not conscious of it so neglect to focus on it in their computer compositions.

Punctuation was found to be erratic in a number of texts from both modes; the most common element was absence of concluding punctuation, the incidence of which was similar for both modes. Another feature was omission of a space between sentences: 8 CB candidates made such errors. It was difficult to determine if the same was true for PB candidates as spacing generally was not uniform.

There were no instances of use of text/email conventions in scripts written in either of the administration modes; this finding is contrary to the researcher's expectations. However, this is encouraging in this more formal setting, though it would be interesting to monitor this over time.

During the analysis one of the most salient points was that errors were more noticeable in the CB scripts; the researcher found it easier to code the transcribed versions of the PB scripts as it was very easy to miss errors in the handwritten scripts. This is an issue for raters, as discussed earlier; training needs to incorporate discussion of this fact and how to address it.

Spelling errors

As stated above the words 'not in the lists' from the Range output contained a mix of proper nouns, topic-specific vocabulary, words that are incorrectly spelt and non-English words. The words that fell into the last two categories were classed as errors. The proportion of error tokens in the 'not in the lists' category was 31% for CB texts and 33% for PB texts. The proportion of error tokens out of the total of all tokens was 3% for both modes. Thus the modes appear to have similar error rates.

However this does not reveal information about the types of error. Four percent of the errors in CB texts could have been the result of typographical errors: for example 'aks' for *ask* and 'tehy' for *they*. Only one possible case of this reversing the order of letters was found for the PB group ('sporst' for *sports*). When word-processing in non-exam situations, spell-checkers would have either flagged up or auto-corrected these errors. For experienced word-processor users there may be an over-reliance on these features that impact on their performance in a test situation. It is unlikely that these would be penalised due to their infrequency but it is worth CB candidates paying extra attention to this when reviewing their work. For errors that impede understanding there were 6% in the CB scripts and 11% in the PB scripts. One reason for the higher number of impeding errors in PB texts was illegible handwriting. These issues have implications for rating as illegibility is only an issue for PB scripts and typographical errors are only an issue for CB scripts.

Conclusion

Returning to the research question, which asked whether written output produced via the two modes of administration was comparable in terms of a number of features, a summary of findings by feature can be presented as follows:

- Lexical resources: texts produced in both modes were of comparable length and demonstrated appropriate vocabulary use; however the CB texts appeared to show more lexical variation.
- Sentence/paragraphing: texts produced in both modes showed differences in both the number of sentences and paragraphs.
- Surface features: texts produced in both modes showed problems with punctuation and capitalisation.
- Errors: lexical error rates in texts produced in both modes were similar although the nature of the errors differed.

This study reveals that informal letter texts composed by hand or by computer under timed assessment conditions do show some variation in linguistic and text features. The findings from this study provide useful insights into the comparability of PB and CB written output within an assessment context. The use of live data show that findings are not an artefact of an artificial or trial situation and thus results have an immediate and direct relevance to testing.

It should be noted that the samples used in the study were limited by those that took the CB test, which currently has a smaller candidature than the PB test. As a result only candidates from three countries were studied. The main L1s used in these countries use a roman alphabet, thus the findings may not necessarily be the same for those using a non-Roman alphabet.

In addition the study involved two separate samples of candidates and while the two groups were matched on language ability it could be that the results were affected by the nature of the samples themselves. The availability of more candidate information such as age, gender and L1 would have lent weight to the matching of the sample. The results from this study have a number of implications for teachers, testers and raters. For teachers, issues of capitalisation, punctuation and paragraphing can be highlighted. If teachers are made aware of areas in which candidates are having problems then these can be addressed at classroom level. For language testers the differences found between the texts written in different modes can be built into rater training. This is both for the interpretation of mark schemes and for dealing with the presentation effect. For example if text organisation in terms of paragraphing is to be assessed then there needs to be an awareness that this may differ depending on the mode in which the text was produced.

This study can be considered a starting point into research into the text differences between PB and CB written output. It would be interesting to explore the effect

found with different L1 groups, those with different writing systems (e.g. Cyrillic/Arabic) and those at different proficiency levels. An exploration of how these features differ across modes in other genres would also be worthwhile. Other text features could also be analysed, for example structure, cohesion and larger lexical chunks. More in-depth exploration of the differences found would give further insights enabling us to see how, for example, the lexical breadth differs between groups, i.e. what types of words are used differently.

Investigation of the writing process using protocol studies would add insight into the composition process and perhaps shed light on why differences and similarities between the texts are found. There would also be value in exploring how the differences found affect the rating process, for example how examiners respond to texts presented differently in terms of paragraphing or the number and length of sentences. An understanding of any effect could feed into rater training to enhance the fairness and reliability of written assessment.

References and further reading

Alderson, J C (2000) *Assessing Reading*, Cambridge: Cambridge University Press.

Bachman, L F and Palmer, A S (1996) *Language Testing in Practice*, Oxford: Oxford University Press.

Goldberg, A, Russell, M and Cook, A (2003) The effect of computers on student writing: a meta-analysis of studies from 1992–2002, *Journal of Technology, Learning and Assessment*, 2/1, retrieved from www.jtla.org

Horkay, N, Bennett, R E, Allen, N, Kaplan, B and Yan, F (2006) Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP, *Journal of Technology, Learning and Assessment*, 5/2, retrieved from www.jtla.org

Lee, H K (2004) A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test, *Assessing Writing*, 9/1, 4–26.

Li, J (2006) The mediation of technology in ESL writing and its implications for writing assessment, *Assessing Writing*, 11/1, 5–21.

MacCann, R, Eastment, B and Pickering S (2002) Responding to free response examination questions: computer versus pen and paper, *British Journal of Educational Technology*, 33/2, 173–188.

Nation, P and Heatley, A (1996) *Range*, Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington, retrieved from www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx

Powers, D, Fowles, M, Farnum, M, and Ramsey, P (1994) Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays, *Journal of Educational Measurement*, 31/3, 220–233.

Russell, M and Plati, T (2002) Does it matter what I write? Comparing performance on paper, computer and portable writing devices, *Current Issues in Education*, 5/4, retrieved from http://cie.ed.asu.edu/volume5/number4

Russell, M and Tao, W (2004a) Effects of handwriting and computer-print on composition scores: a follow-up to Powers, Fowles, Farnum and Ramsey, *Practical Assessment and Research Evaluation* 9/1, retrieved from http://pareonline.net/getvn.asp?v=9&n=1

— (2004b) The influence of computer-print on rater scores, *Practical Assessment and Research Evaluation* 9/10, retrieved from http://pareonline.net/getvn.asp?v=9&n=10

Scott, M (1998) *WordSmith Tools*, Version 3.0, Oxford: Oxford University Press.

Warschauer, M (2003) Technology and second language writing: Researching a moving target, in Kei Matsuda, P, Canagarajah, A S, Harklau, L, Hyland, K and Warschauer, M, Changing currents in second language writing research: A colloquium, in *Journal of Second Language Writing,* 12/2, 151–179.

Weigle, S C (2002) *Assessing Writing*, Cambridge: Cambridge University Press.

Yates, S J (1996) Oral and written linguistic aspects of computer conferencing: A corpus-based study, in Herring, S C (Ed.) *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*, Philadelphia: John Benjamins Publishing Co.

Assessment for teaching: Cambridge ESOL's CLIL exam

MICK ASHTON ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

EVELINA GALACZI RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

Introduction

Since its introduction in 2005, the Teaching Knowledge Test (TKT) has proved a popular choice for governments and institutions looking for an accessible and flexible way to assess teachers' knowledge about teaching . Those involved in developing TKT were hopeful that the provision of a test such as TKT would result in the development of a wide range of courses, which in turn would have the positive washback effect of providing greater numbers of teachers worldwide with opportunities for professional development (see Harrison 2007 for more details of TKT adoptions and courses). Since 2005, more than 90,000 module entries have been made by teachers in over 70 countries and Gerardo Valazza's article in this issue on the impact of TKT in Uruguay further confirms the successful achievement of its objectives (Valazza 2008).

This article outlines the background to the launch of the new TKT: CLIL module in October 2008. In it, we consider the growth in practice of CLIL and the need to expand the TKT framework with a specialist module on CLIL, define the CLIL construct and the testing focuses, and highlight the quality assurance procedures and findings from the main trialling exercise.

What is CLIL?

CLIL operates in a range of contexts and is therefore open to different interpretations. A quote from Marsh helps to explain what CLIL is: 'CLIL is defined as an approach in which a foreign language is used as a tool in the learning of a non-language subject in which both language and the subject have a joint role' (Marsh cited in Coyle 2006:1).

There are many different types of CLIL programmes, ranging from full immersion (Canada) through partial immersion, about 50–60% of the curriculum (parts of Spain), to language showers, regular, 20–30 minute subject lessons in the target language (parts of Germany). In Secondary schools, subjects are usually taught in the target language by non-native speaker subject teachers. In Primary contexts, CLIL programmes are commonly delivered by non-native subject teachers or by English language teachers. In some countries native speaker classroom assistants support the learners too. There are also contexts where native speakers teach English to non-native learners (often from minority language groups) to enable them to integrate into mainstream classes. Examples of these programmes are EAL (English as an Additional Language) in Britain and CBI (Content Based Instruction) in the US.

The Eurydice Survey (2006) provides a detailed review of the provision of CLIL across Europe.

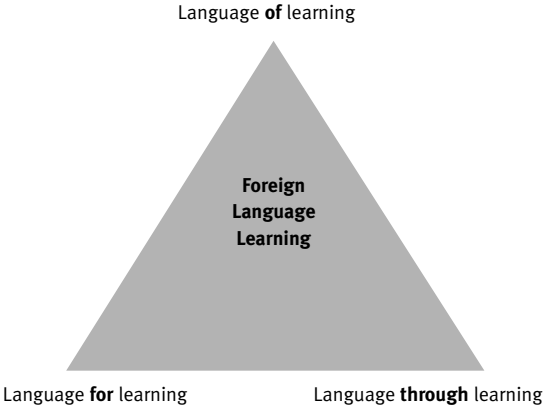
CLIL syllabuses

Traditional EFL syllabuses have been based on grammatical structures, with the order in which the language items are presented being determined by how difficult these linguistic forms are believed to be. When a subject is taught through a second language this changes completely and language takes on a different role. In CLIL contexts interaction in learning is important because students need to use and develop language:

- of learning (the content)
- for learning (interaction – pair work 'think, pair, share' and group work)
- through learning (for cognitive skills).

CLIL programmes require teachers and course designers to reconceptualise their view on language learning, as Figure 1 shows.

Figure 1: Reconceptualising language learning – a language-using model



Coyle (2007:552) sums up this reconceptualisation in the following way: 'Applying this triptych linguistic approach marks a shift in emphasis from language learning based on linguistic form and grammatical progression to a more 'language-using' one which takes account of functional and cultural imperatives.' In other words, as Marsh and Langé (2000) state in the title of their seminal article, CLIL involves 'learning to use language and using language to learn'.

The main aims of CLIL are to develop:

- Content – subject knowledge and skills
- Communication – meaningful interaction
- Cognition – thinking skills
- Culture – an awareness of other cultures, of what it is to be a responsible citizen, and of being part of local and global communities.

These aims are commonly referred to as the '4 Cs'. The common *content areas* include: Art, Citizenship, Classics, Design Technology, Economics, Environmental Studies, Geography, History, Information Computer Technology (ICT), Literacy, Maths, Music, Physical Education (PE), Philosophy, Politics, Religious Education (RE), Science, Social Science. The European Commission for Education and Culture states that CLIL should develop effective *communication skills* for real practical purposes. This is achieved by maximising student talking time in the target language. Learners need to develop *cognitive academic language proficiency* from a young age as this is a prerequisite for studying curriculum subjects. The aforementioned Eurydice Survey refers to the importance of 'conveying to pupils values of tolerance and respect towards other *cultures* through the use of the target language.' (Eurydice 2006:22). So CLIL aims to help learners to build intercultural knowledge and understanding.

The '4 Cs' is not a theory, but a conceptualisation of CLIL that is rooted in a philosophical stance with education being in prime position. It views education in a broad sense, being holistic and inclusive rather than a set of individual disciplines. The '4 Cs' framework 'espouses socio-cultural theory where the social construction of knowledge and culturally embedded learning permeate the whole' (Coyle 2007:552).

Is CLIL successful?

There are many advantages to the CLIL approach: it develops confident learners and enhances academic cognitive processes and communication skills. CLIL encourages intercultural understanding and community values. In addition, research led by Johnstone and McKinstry (2008) shows that learners become more sensitive to vocabulary and ideas presented in their first language as well as in the target language and as a result they gain more extensive and varied vocabulary. In the target language, learners reach proficiency levels in all four skills of listening, speaking, reading and writing far beyond what is expected in TEFL programmes. This success is shown in ICT skills too. Along similar lines, Coleman (2006:37) reports on research in Dutch Secondary schools which indicates that, 'CLIL leads to better English proficiency, that it has no negative effect on Dutch [L1] proficiency, nor on the pupils' subject knowledge' and Van de Craen *et al* (2008:73) conclude that 'CLIL

induces the learner to be more cognitively active during the learning process'.

Why has CLIL grown?

There have been many factors that have contributed to the growth of CLIL. We have already discussed the cognitive and cultural focuses that are commonplace within a CLIL curriculum, and would now like to highlight several other key drivers that have been influential.

There have been socio-political pressures, e.g. European legislation dating back to the 1995 White Paper, *Teaching and learning: Towards the Learning Society*, and policies such as *A New Framework Strategy for Multilingualism* (European Commission 1995, 2005). There is a drive in European political circles towards achieving a greater degree of plurilingualism and making Europe the most competitive and knowledge-based economy in the world. An ambitious undertaking has been set up to enable all Europeans to communicate in two community languages in addition to their mother tongue. See De Graaf *et al* (2007:603–4) for a useful summary of recent educational developments in Europe.

Economic motives have also been present. Some schools are given more government money if they offer integrated content and language programmes, for example. Social pressures are evident, in that parents increasingly want their children to be educated in two or more languages as they see this as 'value-added' schooling.

A theoretical rationale for CLIL is given by Williams and Burden (1997) in their definition of the theory of 'constructivism'. This incorporates active participation, learner-centred tasks, experiential learning and a process approach to writing. Also, CLIL practitioners have cited Krashen's (1985) input hypothesis as being widely applicable to the CLIL approach – more targeted, therefore comprehensible, *i + 1* occurs when the target language is driven by subject content since the meaning of the new input language is wholly contextualised.

There are, in addition, clear educational benefits since CLIL provides greater opportunities for exposure to foreign languages, content-based tasks are more interactive, and authentic materials in multiple languages enhance subject inquiry. This in turn is motivational – many students and teachers want and like the challenge.

Testing teachers' knowledge of CLIL

Before we report on the support for the CLIL module through construct validation and our findings from the main trialling exercise, the reader might wish to consider the final CLIL test syllabus and testing focuses. TKT: CLIL is an optional extension module of the Teaching Knowledge Test (TKT). It tests the candidate's knowledge of Content and Language Integrated Learning and concepts related to a CLIL approach. It tests knowledge about subject teaching in a target language and the learning, thinking and language skills which are developed across different curriculum subjects. It tests knowledge of how to plan lessons as well as knowledge of activities and resources used to support a CLIL approach. It also tests knowledge of teaching strategies and how assessment is carried out in CLIL contexts. The test has been divided into two parts with the first 25 items focusing on the candidates' knowledge of CLIL and the principles that underpin effective CLIL practice. The 55 items in Part 2 focus on different aspects of planning, teaching and assessing within CLIL contexts.

The CLIL module is designed to offer maximum flexibility and accessibility for candidates and therefore does not test subject-specific knowledge nor include compulsory teaching practice. However, it is likely that centres and other institutions will offer courses for TKT: CLIL preparation which may include practical issues arising from specific CLIL contexts and some teaching practice.

The CLIL module is intended to be a platform for professional development. As a result of the global need for language learning, particularly for English, candidates who are teachers of other curriculum subjects as well as candidates who are language teachers can add TKT: CLIL to their existing qualifications. This will demonstrate their understanding of how to teach a broader range of subjects for the 21st century. Table 1 shows the structure and focus of TKT: CLIL.

Table 1: Structure and focus of TKT: CLIL

Part	Items	Title	Areas of knowledge
1	25	**Knowledge of CLIL and Principles of CLIL**	• aims of and rationale for CLIL • language across the curriculum • communication skills across the curriculum • cognitive skills across the curriculum • learning skills across the curriculum
2	(55)	**Planning, Teaching and Assessing**	
	25	2A Lesson Preparation	• planning a lesson and a series of lessons • language demands of subject content and accompanying tasks • resources, including multi-media and visual organisers • materials selection and adaptation • activity types
	20	2B Lesson Delivery	• classroom language • scaffolding content and language learning • methods to help students develop learning strategies • consolidating learning and differentiation
	10	2C Assessment	• focus: content; content and language;cognitive, learning and communication skills • types of assessment • support strategies

One task from the sample paper is shown in Figure 2 which focuses on *learning skills across the curriculum*.[1]

Who is TKT: CLIL for?

TKT: Content and Language Integrated Learning is suitable for pre or in-service teachers teaching across all sectors of education: primary, secondary and tertiary. The CLIL module can be taken at any stage in a teacher's career. It will be of particular interest to:

1. The Sample Paper, along with the Handbook for teachers and a glossary of terms used in CLIL, can be found at www.cambridgeesol.org/clil

Figure 2: TKT: CLIL task on learning skills

For questions **20–25**, match the instructions with the learning skills, listed **A**, **B** or **C**.

Mark the correct letter (**A**, **B** or **C**) on your answer sheet.

Learning skills

A locating information
B organising information
C communicating information

Instructions

20 Listen to the music then beat out the rhythm to your partner.
21 Find three websites to help with your project on recycling rubbish.
22 Decide which information to put on the x axis of the graph.
23 Put the pictures in order of the life cycle.
24 Look in the library books and find two birds that can't fly.
25 Show your picture and explain the symbols in it.

- content teachers who need to teach their subjects in English, or who want to add language teaching to their portfolio of skills
- English language teachers who are required to teach curricular subjects in a second (or third/fourth etc) language
- classroom assistants working in CLIL contexts
- English as an Additional Language (EAL) teachers who have non-native speaker students in their classes.

Construct validation

Grossman's (1995) categories of teaching knowledge underlie the construct of TKT, so they also provide an appropriate framework upon which to map the key areas that a CLIL module should address.

These domains are:

- Knowledge of content
- Knowledge of learners and learning
- Knowledge of pedagogy
- Knowledge of curriculum
- Knowledge of context
- Knowledge of self.

Whilst many teaching decisions are made in accordance with the teaching context, this category of knowledge is not included in the TKT suite due to the difficulties inherent in focusing on local contextual factors in an international test. Professional development, self awareness and reflection in action are all central to teachers in any context. The extent to which a given individual is developing the ability to become an effective reflective practitioner (i.e., 'knowledge of self' above) is clearly beyond the scope of an objective-format test of knowledge.

TKT candidates are encouraged to make use of the Cambridge ESOL Teacher Portfolio, which enables them to keep electronic records of their development as teachers. This does not form part of the assessment of TKT, and as such is an optional rather than a necessary part of TKT.

The content of the TKT: CLIL module corresponds well with the remaining four areas that make up Grossman's framework, and we could map the testing syllabus of the CLIL module on to the Grossman categorisation as shown in Table 2.

Table 2: Comparison of TKT: CLIL testing focuses with Grossman's teaching knowledge framework

TKT: CLIL Part	Categories of teaching knowledge	Example CLIL testing focus
Part 1	Knowledge of content	Language across the curriculum.
	Knowledge of curriculum	Aims of and rationale for CLIL.
	Knowledge of learners and learning	Cognitive, learning and communication skills across the curriculum.
Part 2A	Knowledge of learners and learning	Planning lessons.
	Knowledge of content	Language demands of subject content.
	Knowledge of curriculum	Visual organisers.
	Knowledge of pedagogy	Planning lessons. Activities and resources. Materials selection and adaptation.
Part 2B	Knowledge of learners and learning	Helping students develop learning strategies. Scaffolding content and language learning.
	Knowledge of pedagogy	Classroom language.
	Knowledge of curriculum	Consolidating learning and differentiation.
Part 2C	Knowledge of content	Assessment focus.
	Knowledge of curriculum	Types of assessment.
	Knowledge of learners and learning	Support strategies.

Validation of the CLIL exam

The validation of the CLIL exam is part of the larger context of Cambridge ESOL's ongoing commitment to fairness and *a priori* validation in assessment, with importance given to validation through the collection of data and the role of research in examination processes. The development of the CLIL module also reflects the concern for accountability in the broader assessment context where examination providers are held responsible providing *evidence* of the fitness for purpose of their tests (Saville 2005).

TKT: CLIL was trialled over a period of three months in 11 countries. The principle aim of the trial was to investigate how well the test performed (as a whole and at the item level) and to gather stakeholder feedback which would inform the subsequent stages of the development of the test. The trial population was composed of 259 teachers from a wide range of first language backgrounds and nationalities (Argentina, Austria, Croatia, Germany, India, Italy, Malaysia, Netherlands, Poland, Spain). The top two first languages were Italian (28%) and Spanish (18%). In line

with the anticipated target candidature, the trial population comprised teachers who have qualifications to teach English as well as qualifications to teach other subjects. There was a range of years of teaching experience within the trial population, with the majority of candidates (91%) having had two or more years of experience. In addition, the trial population comprised teachers from both the private and state sector, at primary, secondary and tertiary level.

The trial participants were asked to complete an 80-item trial version of TKT: CLIL, as well as a 25-item language test covering CEFR levels A2 to C2. The language test was used to enable an investigation of the extent to which candidate performance on the CLIL module might be affected by English language proficiency. Background information sheets and feedback questionnaires were also administered to all trial participants and exam administrators in order to gather feedback about different aspects of the examination.

The data analysis encompassed both quantitative and qualitative methods. The quantitative methodologies included item-level Classical and Rasch analyses of item difficulty and discrimination. The qualitative methodologies included questionnaire responses collected after the completion of the trial test.

Findings

We report on three areas below: the performance of the items in the text; the impact of various background variables on test performance and candidates' feedback on TKT: CLIL.

What is the performance of the items on the CLIL exam?

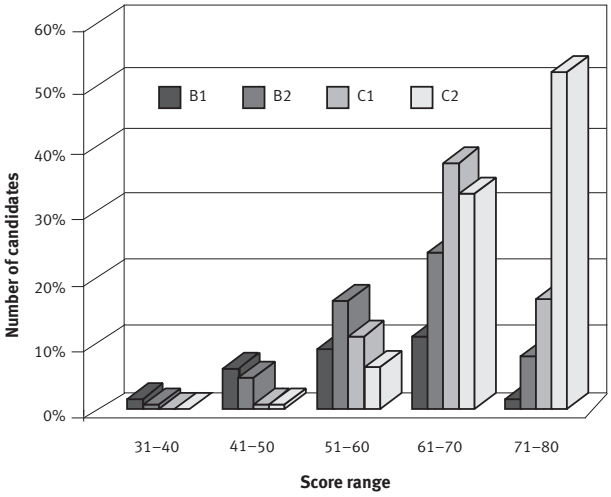
The trialled version of the exam achieved a high reliability of 0.90 (Cronbach's alpha) and a mean P (average facility) of 0.80. Cronbach's alpha (which ranges from 0.0 to 1.0) is a measurement of how consistent the exam items are, for example, in awarding similar marks to candidates of similar abilities. The majority of items had a point biserial value (indicating the discriminatory power of the items) of 0.30 or above. These test item indices provide encouraging validity evidence for the reliability and discriminatory power of the CLIL module as a whole.

What is the impact of background variables on test performance?

One of the research questions investigated whether there was a significant difference in scores between candidates with different levels of *English language proficiency*. As discussed earlier, TKT: CLIL is not a test of English proficiency, but of general CLIL principles and related content areas. Since the test is delivered in English (at CEFR B1 level), however, we need to investigate the relationship between the candidates' English language proficiency and performance on the CLIL module, to make sure the exam does not unduly disadvantage candidates at the lower end of the English proficiency spectrum. A strong effect of English proficiency on CLIL performance could constitute a validity threat to the exam.

The results indicated that in general, the mean CLIL score increased as language ability increased. This is not surprising, since candidates with higher English language ability can understand the questions more accurately, and can process the instructions and items faster. Language proficiency, however, was not found to be an impeding factor, as evidenced by the frequency of marks of candidates in the different language ability groups (see Figure 3). This finding indicates that higher proficiency in English does not automatically lead to a high CLIL score. Alternatively, it also indicates that lower language proficiency is not a barrier to a candidate performing at a satisfactory level on the CLIL exam. It is important to highlight that the relationship between English language proficiency and performance on the CLIL module is not a straightforward one, as teaching experience and, of course, knowledge of the syllabus areas, interact alongside English language proficiency in producing a final score. Due to the relatively small size of the trial population, the different variables could not be controlled, but as the exam goes live this relationship will be further explored and monitored.

Figure 3: Score on CLIL module and English language proficiency



It is also worth noting that the candidate feedback responses indicated that the majority of the teachers in the trial (94%) felt that the language of the test was at the right level.

The findings further indicated that in terms of *teaching qualifications*, candidates who had qualifications both in English and other subjects scored higher (see Figure 4). This finding lends support to the construct validity of the test, as it would be expected that knowledge of teaching both a second language and a subject area would allow candidates to perform better.

Figure 4: Score on CLIL module and teaching qualifications

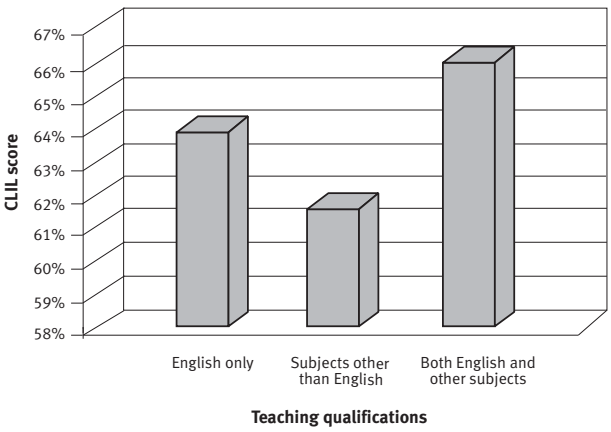
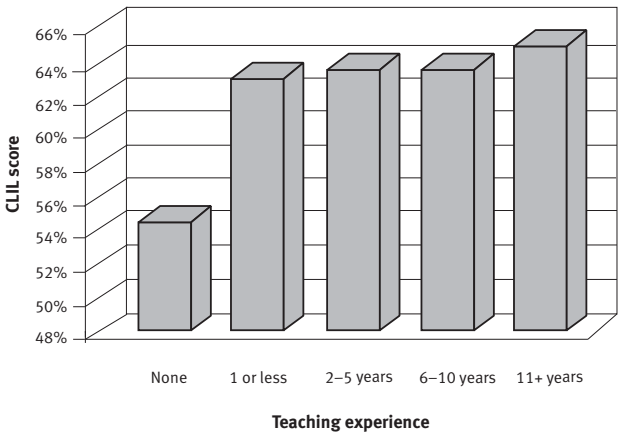


Figure 5: Score on CLIL module and teaching experience



The results also signalled an increase in TKT: CLIL mean score as years of *teaching experience* increased (see Figure 5). This finding can be taken as providing construct validity for the test, as the construct underlying the CLIL exam includes awareness and familiarity with lesson planning, lesson management, and assessment, which would naturally develop as teaching experience increases.

Candidate feedback

Feedback from the trialling has played an important role in the development of the CLIL module. Positive feedback was received in terms of the role of the CLIL module in professional development, and the content coverage of the test:

> It would have been useful for us as teachers if during the last year of our degree, we had had the chance to sit a test like this.
>
> I think I'm not used to thinking about my teaching practice and in one way or another the test has helped me to be more aware of what I do or should do...
>
> The test wasn't exactly enjoyable but very useful for thinking about items I had never done before.
>
> The test was quite comprehensive. Through this test we were able to evaluate and assess out teaching strategies and methodologies.

The candidate feedback also indicated some areas needing further consideration, such as the use of unfamiliar terminology and the need for support materials:

> I would like to say that to do an exam like the second part you must be familiar with the terminology that is used. I have to say that I use some of these strategies and I don't know what they are called.
>
> Before the test is taken by candidates some material for preparation should be made available for candidates.

These valid concerns have been addressed through the TKT: CLIL Glossary of terms and support materials such as the CLIL Handbook and sample paper.

The feedback questionnaire also aimed to investigate candidate perceptions about the authenticity and face validity of the CLIL module. Authenticity was operationally defined as the type of teaching situations candidates and their colleagues may be expected to encounter in a work setting. The results indicated a high level of familiarity with the situations in the test, as 70% of the trial participants responded that they were familiar with the teaching situations encountered in the test. As one trial participant said, "It was good, quite analytical and had many situations that we generally are faced with in our classroom teaching...".

Conclusion

This article has summarised the background to the launch of the TKT: CLIL exam by outlining the basic premises of CLIL and the growth of this new field. It has also described the design, development and quality procedures supporting the CLIL module. As with other Cambridge ESOL exams, this exam has become part of a programme of ongoing research and validation activities which will ensure that satisfactory standards are met in terms of its validity, reliability, impact and practicality.

References and further reading

Coleman, L (2006) CLIL behind the dykes: the Dutch Bilingual Model, Young Learners and the Content and Language Integrated Learning (CLIL) Continuum, *IATEFL YLSIG Online Journal*, 35–40.

Coyle, D (2006) *Developing CLIL: Towards a Theory of Practice, A Monograph*.

— (2007) Content and Language Integrated Learning: Towards a Connected Research Agenda for CLIL Pedagogies, in *International Journal of Bilingual Education and Bilingualism*, 10/5.

Cummins, J (1984) *Bilingualism and Special Education: Issues in Assessment and Pedagogy*, Clevedon: Multilingual Matters.

De Graaf, R, Koopman, G J, Anikina, Y and Westhoff, G (2007) An Observation Tool for Effective L2 Pedagogy in Content and Language Integrated Learning (CLIL), *International Journal of Bilingual Education and Bilingualism*, 10/5.

Eurydice Survey (2006) *Content and Language Learning (CLIL) at School in Europe*, Brussels: Eurydice European Unit.

Fisher, R (2005) *Teaching Children to Learn*, Cheltenham: Nelson Thornes.

Gibbons, P (2002) *Scaffolding Language, Scaffolding Learning*, Portsmouth, NH: Heinemann.

Grossman, P L (1995) Teachers' knowledge, in Anderson, L W (Ed.) *International Encyclopaedia of teaching and teacher education* (2nd ed), New York: Pergamon, 20–24.

Harrison, C (2007) Teaching Knowledge Test Update – adoptions and courses, *Research Notes* 29, 30–32.

Hyland, K (2004) *Genre and Second Language Writing*, Michigan: University of Michigan Press.

Johnstone, R and McKinstry, R (2008) *Evaluation of EPPI: Early Primary Partial Immersion in French at Walker Road Primary School, Aberdeen. Final Report*, retrieved from www.scilt.stir.ac.uk/projects/evaluationwr/documents/EPPI_BOOK.pdf

Jones, G (2001) Bilingual Education and Syllabus design: towards a workable blueprint, in Hall, D (Ed.) *Innovations in Language Teaching*, London and New York: Routledge.

Krashen, S D (1985) *The Input Hypothesis*, London: Longman.

Lasagabaster, D (2008) Foreign Language Competence in Content and Language Integrated Learning, *The Open Applied Linguistics Journal*, 1, 31–42

Marsh, D and Langé, G (2000) *Using Languages to Learn and Learning to Use Languages*, Jyväskalä: University of Jyväskalä, Finland: UniCOM.

Marsh, D (2002) (Ed.) *CLIL/EMILE – The European Dimension: Actions, Trends and Foresight Potential. Public Services Contract DG EAC*, Strasbourg: European Commission.

Marzano, R (2001) *Classroom Instruction that works*, USA: ASCD.

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

Saville, N (2003) The process of test development and revision within UCLES EFL, in Weir, C J and Milanovic, M (Eds), *Continuity and Innovation: Revising the Cambridge Proficiency in English examination 1913–2002*, Studies in Language Testing 15, Cambridge: UCLES/Cambridge University Press, 57–120.

Valazza, G (2008) Impact of TKT on language teachers and schools in Uruguay, *Research Notes* 34, 21–26.

Van de Craen, P, Mondt, K, Allain, L and Gao, Y (2008) Why and How CLIL Works, *Vienna English Working Papers (Views)*, 16/3, 70–78.

Williams, M and Burden, R (1997) *Psychology for language teachers*, Cambridge: Cambridge University Press.

Impact of TKT on language teachers and schools in Uruguay

GERARDO VALAZZA HEAD, ACADEMIC DEPARTMENT, INSTITUTO CULTURAL ANGLO URUGUAYO, MONTEVIDEO

Introduction

The Cambridge ESOL Teaching Knowledge Test (TKT) was administered in Uruguay for the first time in 2005. Since then, candidature in Uruguay has increased significantly, with more than twelve hundred module entries since its first administration. With a country population of 3.3 million (Instituto Nacional de Estadística 2007:10) this means that a TKT module has been taken by around one in every 2,600 inhabitants.

This rapid increase in TKT candidature motivated *Instituto Cultural Anglo Uruguayo* (the Anglo) to conduct the present research study. The Anglo is a private language institute founded in 1934, with 47 branches throughout the country. It has been a Cambridge ESOL authorised centre since 1948 and is currently the supervising centre in Uruguay, where there are four other TKT sub-centres.

The main purpose of this study is to explore stakeholders' perceptions of TKT with a view to answering two main questions:

- What impact has TKT had so far on English language teachers and language schools in Uruguay?
- Can TKT candidature in Uruguay be expected to increase at the same rate in the near future?

It is hoped that the present research study can contribute to our understanding of the impact that assessment can have upon learning, more specifically upon *teacher* learning. In addition, it is expected that the results reported in this paper will help to explain the development of TKT candidature in Uruguay and will provide TKT trainers and language institutions running TKT courses in this country with relevant and useful information which can be of help when evaluating and updating their respective TKT courses.

Main characteristics of the research study

According to Freeman (1996:360), since the early 1980s, research into teacher learning has experienced a methodological shift from the *process-product research paradigm*, which regards teaching as 'external assessable behaviour' on the basis of student learning and achievement, to the *hermeneutic research paradigm*, which aims to 'understand and interpret actions from the perspective of participants,' i.e. from the perspective of the teachers themselves. It is within the latter methodological paradigm that the present study can be framed.

On a more general note, this piece of research can be described as *interpretive* or *naturalistic*. This means it regards theory as 'emergent,' i.e. following research rather than preceding it, it is characterised by a 'concern for the individual' and it aims to 'understand the subjective world of human experience' (Cohen, Manion and Morrison 2000:22–3). In other words, the results are not necessarily generalisable to other contexts. On the contrary, the findings probably highlight the uniqueness and idiosyncrasy of the Uruguayan reality.

Research design

Data were collected by means of the administration of questionnaires in February 2008. Most of the questions were of a closed nature requiring the respondent to choose one or more among a number given options. However, where appropriate, these were followed by open questions in each section in order to enable the respondent to add anything that may not have been contemplated by the researcher.

It should be noted that many of the questions, especially those which enquired about the impact of TKT, required the respondent to grade perceived impact on a Likert scale. Figure 1 shows a sample question taken from the TKT candidate questionnaire.

Validity and reliability of the research study

Two different questionnaires were designed to collect information from two types of stakeholders: TKT candidates and organisations which have run courses leading to this teacher award. The information gathered from these two sources was compared and contrasted with a view to increasing the validity of the findings of the present research study.

As far as the reliability of the research study is concerned, this is not understood as the comparability or transferability

Figure 1: Sample question from TKT candidate questionnaire

11. How far has preparing for and/or sitting for TKT helped you to develop as a professional teacher? Please tick as appropriate using the following scale:

3 = Very much	2 = To some extent	1 = A little	0 = No

	3	2	1	0	Don't know
a) I have developed a better understanding of the theoretical principles underlying my teaching					
b) Now I reflect on my teaching more than I used to					

of the results but as the 'fit between what researchers record as data and what actually occurs in the natural setting that is being researched, i.e. the degree of accuracy and comprehensiveness of coverage' (Cohen *et al* 2000:119).

These issues of accuracy and comprehensiveness of coverage were addressed through the use of electronic mail. I considered this to be the best means of administering the questionnaires because it enabled me to reach:

- stakeholders all over the country very quickly and at a very low cost
- stakeholders who I had previously been unaware of but who had attended different teacher training and development events at the Anglo and as a result had become part of the institute's external teachers database
- not only stakeholders connected to the Anglo but also stakeholders linked to the four TKT sub-centres, thus giving us a more comprehensive view about the positioning of TKT in Uruguay.

In other words, it is believed that the use of email for questionnaire administration purposes actually served to increase both the validity and reliability of the present research study.

Threats to the validity and reliability of the research study

The TKT candidates' questionnaire was returned by 72 teachers who had sat for a total of 181 modules out of the 1,242 modules administered in Uruguay up to December 2007. This figure constitutes 15% of the total TKT candidature measured in modules. A question that remains to be answered is how representative of the whole TKT candidature population the research sample may be.

In order to attempt to answer this question, I analysed the distribution of candidates per module in the research sample and the whole TKT population in Uruguay. Two observations can be made. In both cases, Module 2 candidature is smaller than Module 1 candidature, and Module 3 candidature is in turn smaller than the candidature for Module 2, secondly, the percentage differences between the candidatures for Modules 2 and 3 in comparison to the Module 1 candidatures for the whole TKT population and the research sample seem to follow a similar pattern (see Table 1).

Table 1: Distribution of candidates per module

Whole TKT candidature		TKT research sample	
Modules taken	**Difference**	**Modules taken**	**Difference**
Module 1: 427		Module 1: 62	
Module 2: 411	-3.7%	Module 2: 60	-3.2%
Module 3: 404	-5.4%	Module 3: 59	-4.8%

In other words, these figures may be an indication that the research sample is, after all, representative of the whole TKT population in Uruguay.

As far as the TKT organisations' research sample is concerned, it is much more difficult to measure how representative of its respective whole population it may be. This is due to the fact that whereas the TKT total candidature is known, the population of organisations offering TKT courses in Uruguay remains uncertain. What can be ascertained is that nine organisations running TKT courses completed and returned the corresponding questionnaires.

However, it is worth restating that the purpose of the present research study is not to attain generalisability or transferability of the results but to probe into stakeholders' perceptions of TKT in order to achieve a better understanding of the impact that TKT has had upon them and to obtain information that helps us to understand and predict the development of TKT candidature in Uruguay.

Analysis and discussion of the data

In this section we cover candidates' expectations, profiles, and reasons for taking TKT; why organisations chose to run TKT courses and the impact of TKT on candidates.

TKT candidates' expectations

When realising that the number of TKT candidates decreases from Module 1 to Module 3, I was concerned that this could be indication of the fact that candidates' expectations were not being fulfilled. For this reason, the questionnaire asked candidates how far *sitting* for TKT, on the one hand, and *preparing* for TKT, on the other, had fulfilled their expectations (see Table 2).

If the 'Very much' and 'To some extent' responses are

Table 2: Fulfilment of TKT candidates' expectations

	Very much	To some extent	Very much + To some extent	A little	Not at all	A little + Not at all
How far has *sitting* for TKT met your initial expectations?	35 66%	14 26%	**92%**	2 4%	2 4%	**8%**
How far has *preparing* for TKT met your initial expectations?	46 67%	19 28%	**94%**	4 6%	0 0%	**6%**

considered together as *positive* responses, it could be argued that both sitting for TKT and preparing for it have met candidates' expectations to a very high degree: 92% and 94% of respondents respectively.

As it will have been perceived, some of the 72 candidates who returned the questionnaire did not answer these questions. This can probably be explained by the fact that some of the respondents had prepared for TKT already but had not taken the exam yet at the time they completed the questionnaire, and others had actually taken the exam but had not attended a preparation course. They had either prepared for TKT by themselves or had taken the exam without any specific preparation.

TKT candidates' profile

Three of the preliminary hypotheses about TKT since it was first launched in Uruguay were that it would probably be most appealing to unqualified practising teachers who may not have received any formal training before; inexperienced teachers who had just initiated themselves into the teaching profession, and teachers mostly working on a private basis at home. The present research study was a very good opportunity to gather data that would help either to confirm or refute these hypotheses. The results were very surprising.

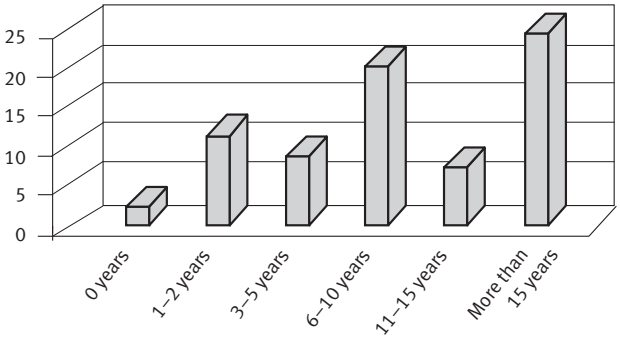
Unqualified vs. qualified

According to the data gathered through the candidates' questionnaire, 34 out of the 72 respondents (47%) had never been formally trained as teachers. The majority, i.e. the remaining 38 teachers (53%), however, had apparently already undergone formal teacher training. Seven of these had completed state teacher training courses and 31 had completed private teacher training courses, nine of which were actually internationally recognised. These figures provide no support for the preliminary hypothesis. On the contrary, the data seem to indicate that *qualified* teachers are apparently finding TKT as appealing as unqualified teachers.

Inexperienced vs. experienced

Thirty-one out of the 72 respondents (43%) had been working as teachers for over 10 years when they sat for TKT, and 24 of these (33%) had been teaching for over 15 years. Figure 2 shows the teaching experience distribution of the TKT candidates who returned the questionnaire.

Figure 2: TKT candidates' teaching experience

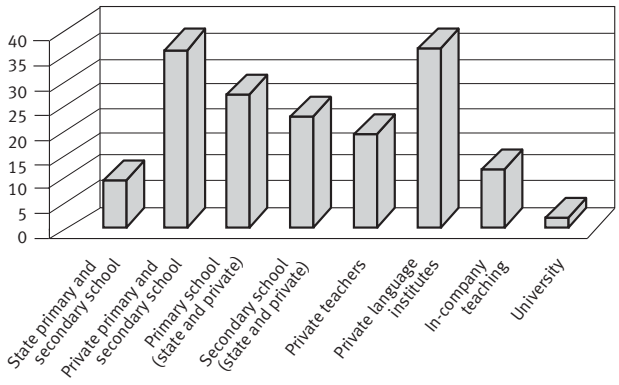


Again, the second hypothesis is refuted by the data. Teachers seem to be interested in TKT regardless of their teaching experience. In fact, the more the experience, the greater the interest in TKT seems to be.

Working context

The third preliminary hypothesis refers to the contexts in which TKT candidates work. Respondents were asked to choose from the following options: state primary school, private primary school, state secondary school, private secondary school, private language institute, in-company teaching, university or privately. Figure 3 shows the results.

Figure 3: TKT candidates' teaching context



The gathered data reveal that although private teachers are in fact sitting for TKT (26%) as anticipated, most of the candidates seem to come from two other sectors: private language institutes (50% of respondents) and private schools (50% of respondents). These results would appear to indicate that the potential for TKT in Uruguay is much greater than anticipated. Why is it that qualified as well as unqualified and experienced as well as inexperienced teachers sit for TKT? Why are teachers from very different contexts finding TKT equally appealing? These are the questions that I attempt to answer in the next section.

Candidates' reasons for taking TKT

In their questionnaire, TKT candidates were asked whhy they chose to do TKT. Being a test of teaching knowledge, it makes sense that the most popular reason was *'To improve my knowledge about teaching'*. This came at the top of the list with 47 out of the 72 respondents (65%) choosing it as one of the reasons why they were doing TKT. Quite expectedly, *'Because TKT is internationally recognised'* appeared high on the list too. It was chosen by 44 out of the 72 respondents (61%).

More surprisingly, 44 teachers also indicated *'To improve my teaching skills'* as one of their purposes for doing the test. It is very interesting to see that even though TKT is a test of *knowledge*, 61% of respondents seemed to expect the test to have an impact on their *teaching practice*. This has important implications for TKT course designers, who may well consider the incorporation of teaching practice or class observation into their TKT courses. This would probably be welcomed by TKT candidates.

Three other purposes indicated by the candidates were: *'To improve my English'*, *'Because I needed formal training'* and *'To get a better job'*, which were pointed out

by 18 (25%), 16 (22%), and eight (11%) candidates respectively. Finally, six respondents (8%) admitted that their employers had asked them to sit for TKT.

In the next section I discuss how the candidates' reasons for doing TKT compare with the organisations' reasons for running TKT courses.

Organisations' reasons for running TKT courses

The information gathered by means of the candidates' questionnaire was compared with data coming from the organisations running TKT courses. Nine organisations returned the corresponding questionnaire and in this section I look into the reasons why these schools and language institutes decided to run TKT courses.

Eight organisations (89%) claimed to have run or be running TKT courses for in-service training and development purposes, which closely matches two of the candidates' main purposes for doing TKT: to improve their knowledge of teaching and develop their teaching skills. In other words, candidates regard TKT as an opportunity for professional development and organisations offer TKT courses to help their teaching staff to develop as professional teachers.

Six out of the nine organisations (67%) reported they had implemented TKT courses to train *external* teachers. In the light of this finding, it is rather strange that only one organisation (11%) admitted to be running TKT because it considered it to be a profitable business opportunity. Why would an organisation train external teachers unless this was considered to be profitable?

Two other reasons indicated by the organisations were the prestige that Cambridge ESOL brings to the institution running TKT, pointed out by five organisations (56%), and the fact that TKT was a good opportunity for teacher training and/or teacher development in areas where there were no other similar options available to teachers. This latter purpose was pointed out by four out of the nine organisations which returned the questionnaire (44%).

The impact of TKT on candidates

Likert scales were used to probe the candidates' perceptions about the impact that TKT may have had on them. Candidates were required to gauge how much preparing for and/or sitting for TKT had helped them to develop as professional teachers with regard to a number of given criteria. The grading options which candidates could choose from were: 'Very much' (3), 'To some extent' (2), 'A little' (1), and 'Not at all' (0). In addition, candidates could opt for 'Don't know' if they were uncertain about the impact of any of the criteria.

Table 3 shows the number of 'Very much' and 'To some extent' responses given for each of the different criteria. In addition, the table indicates the corresponding percentages out of the total number of returned questionnaires, i.e. 72. The last column in the table shows the percentages of these two answers grouped together. For analysis purposes, it is believed that the 'Very much' and 'To some extent' answers can be considered together as an indication of *positive* impact.

A number of important observations can be made from the previous analysis of data. In the first place, with the exception of the last two items on the list, all the criteria were very highly assessed, receiving between 72% and 86% positive responses. As a result, it could be argued that TKT seems to have had an overall positive impact on candidates. Secondly, TKT seems to have had a positive

Table 3: Impact of TKT upon candidates as perceived by themselves

	Very much		To some extent		Very much + To some extent
I have developed a better understanding of theory	34	47%	28	39%	86%
I feel more confident to take part in discussions about teaching and learning	34	47%	27	38%	85%
I developed greater self-awareness of positive aspects of my teaching	32	44%	26	36%	81%
I developed greater self-awareness of negative aspects of my teaching	29	40%	29	40%	81%
I feel more confident about my teaching skills	34	47%	22	31%	78%
My teaching skills have improved	25	35%	31	43%	78%
I have more ideas for the classroom	27	38%	28	39%	76%
I feel better equipped to evaluate teaching materials	24	33%	31	43%	76%
TKT has motivated me to continue developing as a professional teacher	35	49%	19	26%	75%
I reflect more on my teaching	25	35%	29	40%	75%
I plan for and meet my students' needs more effectively	25	35%	29	40%	75%
My students can learn more from my lessons	25	35%	29	40%	75%
I can identify my students' needs more precisely	22	31%	32	44%	75%
I find teaching more enjoyable	21	29%	33	46%	75%
I can understand articles and books on TEFL more easily	21	29%	31	43%	72%
I plan my lessons more thoroughly	24	33%	25	35%	68%
My students enjoy my lessons more	10	14%	30	42%	56%

impact on candidates not only from a *theoretical* point of view but also from a *practical* perspective. Some of the criteria on the list, for example, 'I have developed a better understanding of theory' and 'I feel more confident to take part in discussions about teaching and learning,' are naturally linked to knowledge of theory. On the other hand, other statements such as 'I have developed greater self-awareness', 'I have improved my teaching skills', 'I feel more confident about my teaching skills', and 'I have more ideas for the classroom' are all connected to teaching practice. Finally, even though 'I plan my lessons more thoroughly' and 'My students enjoy my lessons more' were not as highly assessed as the other criteria, the scores they received were still very interesting: 68% and 56% respectively. It should be noted that some candidates did not choose 'Very much' or 'To some extent' for the former criterion because they believed they were already planning lessons very thoroughly before doing TKT, and perhaps the latter criterion was not assessed more highly due to the intrinsic difficulty in assessing students' enjoyment in class. In the next section I analyse the impact of TKT from the point of view of the organisations which have run TKT courses in Uruguay.

The impact of TKT on organisations

The same Likert scales and criteria described in the previous section were adapted and included in the questionnaire administered to the organisations which have run TKT preparation courses in order to look into their perspective of the benefits derived by the course participants. Nine organisations returned the questionnaire – see Table 4 for the results.

The first five criteria on the table were very highly assessed, and the following five received a score of 67%. Three of the remaining criteria were rated at 56% and the last two items were rated at 33%. Although these results are not as high as the results obtained through the candidates' questionnaire, it could still be argued that the organisations in question believe that the overall impact of the TKT courses upon the participants has been positive.

In addition to this, it should be noted that out of the first five criteria rated most positively by the organisations, two were also very highly rated by the candidates themselves. Firstly, the fact that the organisations believe that TKT course participants have improved their knowledge about teaching (89%) is related to the candidates' perception that they have developed a better understanding of theory, which came up at the top of the list for them, rated positively by 86% of the candidates. Secondly, the fact that according to the organisations there seem to be more discussions about teaching in their staff rooms (78%) may be a consequence of the teachers feeling more confident to discuss aspects of teaching and learning, which was rated positively by 85% of the candidates.

These two criteria highlight the impact that TKT has had upon course participants' *knowledge about theory*. In addition, the impact of TKT upon the course participants' *teaching practice* is also highlighted by the organisations. They have rated positively the criteria indicating that course participants seem to have become more aware of their students' needs and interests (89%), have more ideas for their lessons (78%), have become more confident in their teaching skills (67%), have developed self-awareness regarding their teaching (67%) and plan their lessons more thoroughly (67%).

Table 4: Impact of TKT upon candidates as perceived by the organisations

	Very much		To some extent		Very much + To some extent
They seem to be motivated to engage in further teacher training/ development activities	5	56%	3	33%	89%
They seem to have become more aware of their students' needs and interests	5	56%	3	33%	89%
They seem to have improved their knowledge about language teaching	4	44%	4	44%	89%
There are more discussions about teaching and learning in the staff room	6	67%	1	11%	78%
They seem to have more ideas for their lessons	4	44%	3	33%	78%
They seem to feel more confident about their teaching knowledge and/or skills	5	56%	1	11%	67%
They seem to be more motivated to do their job	5	56%	1	11%	67%
They seem to have become more aware of positive and/or negative aspects of their teaching	4	44%	2	22%	67%
When they discuss aspects of teaching they refer to the materials they have read during the course	3	33%	3	33%	67%
They seem to plan their lessons more thoroughly	2	22%	4	44%	67%
They seem to have become more aware of positive and/or negative aspects of the materials they use	4	44%	1	11%	56%
Their students have obtained better results at the end of the course	2	22%	3	33%	56%
They seem to have improved their teaching skills	2	22%	3	33%	56%
Their students have made positive comments about them and/or their teaching	2	22%	1	11%	33%
Absenteeism in their groups has decreased	1	11%	2	22%	33%

In conclusion, it could be argued that the overall impact of TKT upon candidates has been positive. This is highlighted by the candidates themselves and seems to be corroborated by the organisations which have run TKT preparation courses.

Other benefits which have been mentioned, highlighted by 56% of the organisations, are that the TKT courses have been financially profitable and that the organisations have become points of reference for other local institutions and teachers. In addition, 33% of the organisations claimed to have earned prestige as a result of running TKT courses and that they have developed a reputation as teacher training centres.

Perceptions over time

I thought it would be interesting to see how perceptions of TKT have developed over time. For this purpose, I looked into what 2005 and 2006 candidates had to say about TKT and the impact it has had upon them. Apparently, opinions about TKT continue to be positive in spite of the passing of time. The following quotes have been taken from some of the open questions in the candidates' questionnaire:

> It (preparing for TKT) was an opportunity to evaluate myself, my teaching, even my goals in the classroom.
>
> TKT is an excellent tool for teachers to assess their knowledge of the theory and practice of teaching languages.
>
> It was useful and refreshening [*sic*] at the time. I think it is good to revise and polish one's knowledge from time to time. What's more it's an opportunity to update professional terminology and acquire new concepts and perspectives.
>
> I strongly believe that the TKT course, because of its flexibility and accessibility, is a great opportunity for those who want to get involved in professional knowledge of the EFL teaching. I personally think that TKT is a real road to teaching where your possibilities widen as you can get a better understanding of the complexities of this profession. To say the last but not the least, TKT is also a challenging road to go on learning.

Impact of TKT on Uruguayan society

The data gathered in the present research study would appear to indicate that TKT has had a very positive impact on TKT candidates and the organisations that have run TKT courses in Uruguay. Could it be argued, as a result, that TKT is having a positive impact on Uruguayan society as a whole?

It is difficult to arrive at such a conclusion from the data which have been collected and the analysis I have presented here. Nonetheless, it could be tentatively argued that if TKT is having such a positive impact upon teachers, their teaching could in turn be having a positive impact on the students' learning of English. If students' learning were being enhanced through TKT, students' chances of furthering their studies and obtaining better jobs might be enhanced too.

However, further research needs to be carried out in order to make any of these arguments more conclusively. The section that follows expands on some of the related investigations which could be made in the future.

Further research

Before any conclusions can be drawn any more persuasively from the present study, it is suggested that further research is conducted. For example:

- TKT candidates could be observed teaching while they are following a TKT preparation course. It would be interesting to find evidence in their teaching of any impact that the course may be having on them.
- TKT candidates' students could be interviewed in order to find out if they have perceived any changes in the work of their teachers while they are attending the TKT preparation course.
- TKT candidates' students' results could be analysed and compared with their students' results in previous years. Are TKT candidates' students doing better as a result of their teachers preparing for TKT?
- Organisations which have run TKT courses could be followed up over time to see if they continue to run TKT courses. Have these organisations' reputation in the local area been enhanced? How have stakeholders' perceptions of the institutions in question changed as a consequence of the latter running TKT courses?

Conclusion

In spite of the caveats expressed in the previous section, I believe that there are sufficient data coming from the questionnaires administered in the present research study to argue that TKT has so far had a positive impact upon candidates and organisations running TKT courses. However, in order to assert that this is also happening at the level of society as a whole in Uruguay, further research is required.

In the light of the results of the present study, I also believe it is possible to argue that TKT candidature in Uruguay can be expected to continue to increase in the near future. The fact that stakeholders' opinions about TKT seem to be so positive supports this view. However, it is more difficult to predict whether TKT candidature will continue to increase at the current high rate.

Finally, it is hoped that the present research study contributes to the positioning of TKT in Uruguay and that even more language institutions and language teachers start to regard TKT as a valid professional development option.

References and further reading

Cohen, L, Manion, L and Morrison, K (2000) *Research Methods in Education*, London: Routledge Falmer.

Freeman, D (1996) The unstudied problem: Research on teacher learning in language teaching, in Freeman, D and Richards, J (Eds) *Teacher learning in language teaching*, Cambridge: Cambridge University Press.

Instituto Nacional de Estadística (2007) *Uruguay en cifras*, retrieved from www.ine.gub.uy/biblioteca/uruguayencifras2007/URUGUAY%20EN%20CIFRAS%2007%20completo.pdf

Impact of the new school-leaving exam of English in Poland

JO LEWKOWICZ INSTITUTE OF APPLIED LINGUISTICS, WARSAW
ELŻBIETA ZAWADOWSKA-KITTEL LINGWISTYCZNA SZKOŁA WYŻSZA, WARSAW

Introduction

Alderson and Wall's (1993) seminal paper *Does Washback Exist?* drew attention to the complexities of washback and highlighted the need for comprehensive washback studies. Since then a number of such studies have been reported in the literature investigating, among others, national language examinations such as those used in Sri Lanka (Wall and Alderson 1993), Hong Kong (Andrews and Fulilove 1994, Cheng 1997) and China (Qi 2004) and international ones such as TOEFL (Alderson and Hamp-Lyons 1996) and TOEFL iBT (Wall and Horak 2008). However, given the particularity of washback and the fact that it depends not only on the nature of the exam but also on the situation in which it is employed (Alderson and Banerjee 2001), washback needs to be investigated for each and every high-stakes examination. This is no less true of the new school leaving examination – Nowa Matura (NM) – introduced in Poland in 2005. Among the compulsory exam papers that matriculating students must pass is a foreign language and the majority of students (around 80% according to the Central Examination Committee [Centralna Komisja Egzaminacyjna]) opt for English. This article thus focuses on the washback of the NM (English). It reports on a questionnaire survey of English language teachers which set out to determine teachers' perceptions of the new examination and how it is impacting on what they teach and how they go about attaining their teaching objectives. Before proceeding to the survey, however, it is necessary to provide some background on the Polish educational system and the nature of the new examination.

Background

A new matriculation examination was proposed in 1999 (Wikipedia) as part of the major reform in education that followed the collapse of Communism in Poland. The new exam was eventually implemented in 2005. Since then, students wishing to matriculate are required to pass examinations in three core subjects, namely Polish, a modern foreign language and an academic subject chosen from the following: biology, chemistry, physics and astronomy, history, geography, knowledge of society and mathematics.[1] In addition, candidates can choose up to three additional subjects not selected as core.

A primary aim of the NM was for all the exams to be standardised and to replace university entrance examinations (Okręgowa Komisja Egzaminacyjna 2004). Until its introduction, although the exams had to follow the basic subject syllabus, the written papers were set at the regional level and hence varied from one region to another and the oral exams were the responsibility of individual teachers. Hence universities felt obliged to test applicants. Under the new system, examination papers are set and marked centrally by trained external examiners, ensuring objectivity of marking as well as comparability of results across the country. However, for language exams, the oral is conducted within the candidate's school and even though neither of the examiners is the candidate's teacher and one of the examiners is appointed from outside the school, universities continue to distrust these and rely on the results of the written exam. To pass, candidates need to attain 30% on each paper (considerably lower than the 50% required under the previous system).

Of relevance to this article is that all candidates are now required to select a modern foreign language and that the exam is made up of two papers: a written test and an oral. Candidates choose between taking both tests at either the basic or extended level or they may opt to take the tests at different levels.

The aim of the language exam is to assess communication and ability to function in the language which is reflected in its structure (Okręgowa Komisja Egzaminacyjna 2004). As can be seen from Table 1 in the written paper at the basic level knowledge of grammar and vocabulary is not assessed separately, while at the extended level it is awarded only 10% of the total score.

At both levels candidates take part in a direct test of oral ability which requires them to interact with an interlocutor (one of the examiners). The focus here is on conveying meaning, with only 5 of the total 20 points being awarded for overall language ability at the basic level and 7 out of 20 at the extended level (see Table 2).

Table 1: Structure of the basic and extended level written exams

Level and timing	Component	Max score
Basic level		**50**
Total time: 120 minutes	Listening comprehension	15
	Reading comprehension	20
	Writing	15
Extended Level		**50**
Part 1: 120 minutes	Use of language (grammar & vocabulary)	5
	Writing	18
Part 2: 70 minutes	Listening comprehension	15
	Reading comprehension & language structure	12

1. The choice of subjects has varied over the years.

Table 2: Structure of the basic and extended level oral exams

Basic level	
Task 1: three guided conversations • asking for/giving information • describing events • negotiating	*Task 2:* Describing a picture and answering two questions related to the picture
Extended Level	
Task 1: Discussion based on stimulus material (such as a picture/ series of pictures and/or graph)	*Task 2:* Presentation and discussion with examiner on the presentation

Teachers' perception of the NM (English)

Since the introduction of the NM (English) there has been much talk among teachers and comment in the press and on the Internet (e.g. Kula 2005, Malicki no date) about problems associated with the new exam. There has, however, been little systematic research investigating any aspect of the exam. A notable exception is the study by Powszytek (2008) which set out to map the NM (English) onto the Common European Framework of Reference (CEFR) levels. Based on an analysis of the 2005 and 2006 papers, he established that the overall level of the basic exam tasks is B1 with individual tasks varying from A2–B2, while that of the extended level is B2 with tasks varying from B2–C1. His study did not, however, address the issue of washback. This study therefore looks at teacher practice and in particular at what teachers do to address the needs of learners preparing for the NM exam. The study was carried out through a detailed questionnaire survey investigating what teachers do in their English language classes and their perceptions of the new exam of English.

Questionnaire survey

An extensive questionnaire based on similar questionnaires (e.g. Qi 2004) and guided by the researchers' knowledge of the context, was developed specifically for this study. It was divided into three parts eliciting background information (Questions 1–11), teachers' classroom practice (Questions 12–18) and teachers' perceptions of the NM (English) exam (Questions 19–29). Most of the questions were closed, though questions 26–29 were open-ended to allow respondents to voice their opinions about the exam. It was first piloted with four practising teachers and their suggestions were incorporated into the final version of the questionnaire which was administered in Polish. The questionnaire was distributed to and with the help of extramural students studying in Warsaw but teaching at various locations throughout the country. This means of distribution ensured that respondents were from various parts of the country and not all located in Warsaw. A total of 57 completed questionnaires were returned and analysed.

Profile of the respondents

At the time of completing the questionnaire, all 57 respondents were teaching students preparing for the NM (English) exam in a public school. In addition, 40 (70%) of the respondents were teaching privately outside the school system. Most respondents were young teachers below 36 years of age (58%, with 11% below the age of 26), and many (47%) had qualified since 2001, at the time leading up to or since the full implementation of the NM exam. All held either the equivalent of a Bachelor of Education degree (37%) or a Masters degree in English philology or linguistics (63%). Respondents' experience in teaching final year students ranged from less than 2 years (14%) to more than 10 years (34%).

The location of the schools in which the respondents were teaching varied with one teacher (2%) teaching in a village school, 14 (25%) teaching in small towns, 17 (30%) in provincial towns, and the remaining 25 (43%) in the metropolis. Most of the schools in which the respondents were teaching had between 500–1,000 pupils (79%) and the majority (75%) had 4 or more parallel classes in each year. The teachers would normally be assigned 2 or more classes (65%) with only 19 (33%) having only one such class.

What is important to note is that many respondents (27 of the 57 who responded) indicated that they taught mixed classes, that is, one or more class where some pupils would be taking the basic level exam while others the extended level. Only 10 of the respondents taught a class specifically designated to take the extended level exam, and even in such cases it would appear that not all the students would sit for the extended level exam. In only two instances did respondents indicate that up to 75% of their students would be taking the extended level exam and none indicated more than 75%. In contrast, many respondents (26) suggested that 10% or fewer of their students would sit for the extended level exam. Many repondents (30 of the 57), thus, were assigned classes at the basic level (some teachers were assigned classes at different levels, hence the discrepancy between the total number of respondents and the above data). The number of hours for preparing students for the exam also varied, from 2 to 4+ per week with the majority of teachers being assigned 3 or 4 hours per week. The number of hours per week seems to depend on individual schools rather than the level of the exam for which the students are preparing.

Achieving teaching objectives

Of interest in the second part of the questionnaire was what teachers teach and how they go about attaining their objectives. The first point to note is that although the amount of time teachers devote to each of the language skills varies from 10–40% of class time for reading, writing and listening and 15–60% for speaking, most respondents devote an approximately equal amount of time per skill. The majority also believe that they create opportunities either frequently or very frequently for practising these skills (see Table 3).

Table 3: Frequency of skills-related class activities

Skill	Never	Rarely	Sometimes	Frequently	Very Frequently
Writing	1 (2%)	11 (20%)	11 (20%)	23 (41%)	10 (18%)
Listening	0	0	9 (16%)	31 (54%)	17 (30%)
Reading	0	1 (2%)	13 (23%)	32 (56%)	11 (19%)
Speaking (role play)	1 (2%)	7 (12%)	11 (19%)	29 (51%)	9 (16%)

To help achieve their objectives, most of the respondents draw heavily on a variety of materials (see Table 4). All rely on practice papers and a majority (58%) use these frequently or very frequently in their teaching. Many (84%) also use exam-type questions frequently or very frequently when teaching the receptive skills. Such questions are likely to be prevalent in the published coursebooks that all but one of the respondents say they employ in their teaching. Most (82%), though not all, draw on the exam syllabus. The majority (88%) also familiarise students with the assessment criteria used during the exam by using these frequently or very frequently for in-class assessment.

Table 4: Materials used in teaching

Materials	Number (%)
Coursebooks	56 (89%)
Practice papers	57 (100%)
Newspapers & magazines	23 (40%)
Internet	37 (65%)
Computer programmes	8 (14%)
TV/radio	11 (19%)
Films, DVDs, video	30 (53%)
Own materials	48 (84%)
Exam syllabus	47 (82%)
Others	3 (5%)

Further evidence that the respondents gear their teaching towards the demands of the NM exam is apparent when the mean ratings for the subskills of writing in terms of importance are compared (see Table 5). What respondents seem to consider most important are aspects of task fulfilment: including in the written text all the points required, adhering closely to the instructions, and understanding the purpose of a given form of writing. This is likely influenced by the fact that at both the basic and extended levels candidates are rewarded for these aspects of writing. On the other hand, respondents appear to value less the mechanics of writing such as punctuation, correct spelling, and grammatical accuracy precisely because these aspects are less heavily weighted in the marking of the exam.

In terms of speaking a similar, though less marked, tendency can be observed (see Table 6): respondents appear to consider the teaching of strategies for the oral test as the most important. They also focus on providing students with opportunities to answer their questions, to participate in pair work, and to make short presentations of the sort that will appear during the exam. They see as less important correcting pronunciation and encouraging grammatical accuracy. They also see understanding culturally related aspects of the language and use of socially-appropriate forms as less important.

Table 5: Importance of writing subskills

Subskill	Importance as mean score (max 5)
Including all points specified in the question	4.84
Adhering closely to the instructions	4.65
Understanding the purpose of a given form of writing	4.64
Linking ideas throughout the text	4.58
Planning the text	4.49
Adhering to the word limit	4.35
Writing a plan	4.18
Addressing the audience	4.05
Using an extensive vocabulary	3.84
Self-correction of written text	3.71
Grammatical accuracy	3.66
Spelling	3.66
Punctuation	3.27

Table 6: Importance of aspects of teaching oral skills

Subskill	Importance as mean score (max 5)
Strategy training	4.54
Providing opportunities for answering Ts questions	4.45
Providing opportunities for pair work	4.42
Providing Ss with vocabulary for description	4.40
Providing Ss with opportunities for making a presentation	4.25
Providing Ss with opportunities for making a presentation of an event	4.23
Providing Ss with opportunities for practising short dialogues	4.11
Introducing visual stimuli for discussion	4.11
Providing Ss with opportunities for group work	4.04
Practising fluency	4.02
Pronunciation	3.95
Working with individual Ss	3.87
Grammatical accuracy	3.79
Correcting Ss grammar	3.65
Introducing/discussing aspects of culture	3.37
Teaching idioms	3.18

The fact that certain aspects of language are considered less important for the exam does not, however, mean that they are not included in the respondents' teaching. Most teachers, for example, spend class time on grammar and vocabulary exercises even though they recognise that only a limited amount of credit is given for grammatical accuracy and use of an extensive vocabulary, especially at the basic level. A total of 47 out of 57 respondents (83%) indicated that they teach grammar and vocabulary at the basic level and 36 out of 43 respondents (84%) do so at the extended level. Many (60%) teach grammar rules explicitly either frequently or very frequently and a similar percentage (54%) incorporate grammar exercises into their teaching frequently or very frequently. Even more respondents (81%) include the specific teaching of vocabulary frequently or very frequently.

Although the majority of respondents tend to limit the amount of L1 (Polish) in their English language classes, most (70%) use it sometimes and as many as 25% use it around 50% of class time. In the open-ended question in Part 3, some of the reasons for using Polish became apparent and these included the limited amount of time to cover the syllabus – using the L1 saves time – and the fact that it is easier to explain grammar and vocabulary as well as test-taking strategies in the L1. Also, all the instructions on the exam papers are in the L1 which was noted as a disadvantage by several respondents, but an advantage by others. This may be the reason why teachers use the L1 for providing instructions.

Teachers' opinions of the NM (English) exam

The focus of the third part of the questionnaire was to ascertain teachers' opinions about the NM (English) exam in terms of its strengths and weaknesses. Questions 21–23 focused on the level of difficulty of the respective papers and it would appear that respondents are divided in their opinions about the basic level exam. Although 46% of respondents consider it to be at the correct level for candidates, 51% consider it too easy. There was greater agreement about the extended level exam with 81% considering it to be of an appropriate difficulty. For many of the respondents (54%), the distance between the two exams is seen as too great and this matter is considered by quite a few respondents (10) as a negative aspect of the NM.

The question of level was raised by a considerable number of respondents. Ten respondents specifically noted that the 30% pass mark is too low and an additional 14 simply reiterated that the exam (basic level) was too easy. The result of such a low pass mark is seen by some to be affecting standards and motivation, as it does not require much effort to get 15 marks out of 50 and as one respondent noted "you have to work hard to get less than 30%".

A related issue raised by three respondents was the difficulty of preparing students for the different exam levels. What appears to happen in at least some instances is that teachers try to address the needs of the majority (i.e. those studying for the basic level exam) and the other students are left to their own devices.

The exam tasks themselves were also seen as problematic by some of the respondents who noted that they failed to promote real communication. This was most often noted in relation to the oral exam at the basic level where candidates work with ready-made scenarios described in Polish. Some felt these encouraged rote learning, provided no opportunity for improvisation and failed to resemble real-life communication. A number of respondents went as far as to suggest that the oral exam should be abandoned given that the results are not taken into consideration by universities. The issues of guessing, getting the right answer for the wrong reason and being able to string words together with little or no grammatical accuracy during the written test were also noted as areas of concern. Some felt that the nature of the exams encourage teaching towards them and two respondents reported that the tests are having a negative effect on their teaching.

However, numerous strengths of the new exam were recorded. The most frequently noted attributes were that the NM has adopted a communicative approach to testing (23 respondents), that the written test is standardised (6 respondents), assessed objectively (18 respondents) by qualified examiners (6 respondents) using clear criteria (12 respondents).

Conclusion

It would appear from this somewhat limited study that the introduction of the NM has had a strong influence on the teaching of English in upper secondary schools in Poland. Respondents, all of whom were teaching NM students, indicated that they focus on teaching communicative skills and tend to stress the need for fluency. They primarily include in their teaching the task types introduced in the exam and they teach test-taking strategies to help students perform well on the exam. Many also provide a considerable amount of test-taking practice and they mark their students' work using the same criteria as on the exam. However, most do not limit themselves to exam practice. They use a variety of materials including films, DVDs and the Internet. They also introduce grammar and vocabulary exercises focusing on accuracy, most probably because they see these as forming the basis of language knowledge. Where they appear to place less stress, however, is on self- and peer correction, thus, they seem to be failing to inculcate into students the need to monitor the language they are using.

Even though the exam seems to be having a strong influence on teaching, this does not mean that the NM has met with the approval of teachers. Many seem to be critical of both the exam format as well as the tasks employed in the exam. The most frequent reservations seem to be levied against the basic level exam, though this may be because it is the level which teachers most frequently deal with. This test is considered by many to be too easy particularly in light of the 30% pass mark which can be obtained with a very limited knowledge of language and which is far below the CEFR standard requirement of 60%. The oral test at this basic level was singled out most frequently as failing to promote the real-life interaction it was meant to encourage (an issue also raised by Powszytek 2008).

The difference between the two levels is also a concern. It may be encouraging students who do not need to take the extended level for university entrance to opt for the easier exam: it is better to do well on the basic level exam than poorly at the extended level. Teachers are thus often faced with mixed classes and the difficulties of addressing the very different needs of their students. The perpetuation of mixed classes we would suggest is one of the most challenging issues facing teachers as a result of the new exam, and one that was certainly not intended when the NM was introduced.

References and further reading

Alderson, J C and Banerjee, J (2001) Language testing and assessment (part 1), *Language Teaching*, 34/4, 213–36.

Alderson, J C and Hamp-Lyons, L (1996) TOEFL preparation courses: A study of washback, *Language Testing*, 13/4, 280–297.

Alderson, J C and Wall, D (1993) Does washback exist? *Applied Linguistics*, 14/2, 115–129.

Andrews, S and Fulilove, J (1994) Assessing spoken English in public examinations – Why and how? In Boyle, J and Falvey, P (Eds) *English language testing in Hong Kong*, Hong Kong: Chinese University Press, 57–85.

Centralna Komisja Egzaminacyjna (2007) *Informator o egzaminie maturalnym od 2008 roku – język angielski*, retrieved from www.cke.edu.pl/images/stories/Inf_mat_od2008/angielski_a.pdf

Cheng, L (1997) How does washback influence teaching? Implications for Hong Kong, *Language and Education*, 38–54.

Kula, M (2005) *Nowej matury nie zdało około 5 proc. Maturzystów* [Five per cent fail the Nowa matura!], Gazeta Wyborcza 27/06/05.

Malicki (no date) *Skołowani-tylko fakty* [Confused – facts only], retrieved from nauka.pl

Okręgowa Komisja Egzaminacyjna, Warszawa (2004), *Informator maturalny od roku 2005 z języka angielskiego* [NM Syllabus for English], Warszawa: Oficyna Wydawniczo-Poligraficzna „ADAM".

Powszytek, P (2008) *European Standards of Testing Language Proficiency and the Polish "Nowa Matura" Examination in English*, Krakow: Universitas.

Qi, L (2004) *The intended washback of the national matriculation English test in China: Intentions and reality*, Beijing: Foreign Language Teaching and Research Press.

Wall, D and Alderson, J C (1993) Examining washback: The Sri Lankan impact study, *Language Testing,* 10/1, 41–69.

Wall, D and Horak, T (2008) *The TOEFL Impact Study: Phase 2. Coping with Change*, TOEFL iBT Research Series No.05, Princeton, NJ: Educational Testing Service.

Wikipedia (2008) Matura, retrieved from http://pl.wikipedia.org/wiki/Program_catereach-reform

Exploring attitudes: the Romanian Bacalaureat versus CAE

CODRUŢA GOŞA UNIVERSITATEA DE VEST TIMIŞOARA, ROMANIA
LUMINIŢA FRENŢIU UNIVERSITATEA DE VEST TIMIŞOARA, ROMANIA

Introduction

This article reports on a questionnaire-based case study which investigates teacher and student attitudes towards two different high-stakes English language examinations popular in Romania: the English tests of the Romanian school-leaving examination, the Bacalaureat, and the Certificate in Advanced English. This study is part of a wider investigation which includes surveys and focus groups run at the secondary and tertiary school level in two towns in Romania. This research will also have a time element, as it will be conducted in two academic years.

The context

To better understand some of the attitudes our respondents put forward, it might be useful to briefly discuss the wider picture of the Romanian educational system and then to focus in on the English component of the Romanian school leaving examination, called the Bacalaureat. As far as the other high-stakes examination under scrutiny is concerned, the Cambridge ESOL Certificate in Advanced English, only its state in Romania will be briefly referred to, since we start from the assumption that those interested in reading this article are already familiar with it.[1]

The totalitarian legacy

Post-communist Romania is undergoing many social changes, in politics, economy and institutions, and is struggling to replace old social values and practices with new ones. It has generally been observed in the Romanian media that a fairly large number of innovations are implemented in a rather haphazardous manner without much preparation or planning. As Fullan (1995:4) argues: 'Neglect of the phenomenology of change – that is, how people actually experience change as distinct from how it might have been intended – is at the heart of the spectacular lack of success of most social reforms.' An illustration of Fullan's argument is the way in which the Ministry of Education has attempted (and the process is still ongoing at this date) to 'reform' the Bacalaureat, an issue that will be explored in more detail later in this article.

Before December 1989, the system of education in Romania was shaped by philosophies common to the totalitarian regimes in Central and Eastern Europe: it was highly centralised, totally subordinated to the ideology of the Communist Party, aimed at creating 'a new person' (completely devoted to the Party ideology: ready to renounce family, property, freedom of thought and expression for the country and Party). Bârzea (1998) observes that the ultimate aim of communist education could be represented as follows: one ideal = one nation = one educational system = a new person. Under the circumstances, teachers were expected to be particularly loyal to this ideology and transmit it to their students. Beresford-Hill (1994:50) presents the real features of 'Homo Sovieticus', the new species produced by the communist regime, as follows: 'a political passivity, a reluctance to accept or to undertake public obligations, a cynical negativism and an almost childlike expectation of instant gratification.'

These features, in many cases, turned out to be for the outside world only. Most people seemed to fake them in public while privately they were looking for means of surviving total immersion in the muddy totalitarian waters. Matei (2002), when discussing the state of affairs in pre-1989 Romania, observes that only in small trusted circles could you allow your real self to surface, because otherwise there was a great risk of the infamous Securitate (as the Romanian Communist Secret Police is known) agents being present and 'telling on you'. Matei (2002:27) observes:

> 'In such small family or friendly communities you would feel safe enough to indulge in large, savoury portions of political jokes about Ceausescu and the 'Commies' – the strongest antidote to your dreary daily existence under the watchful eye of the Big Brother.'

1. For further information about CAE visit www.CambridgeESOL.org/exams

To sum up, during the communist regime in Romania, everyone, but most of all teachers, had to display total adherence and fervent support to what was called 'political-ideological education'. That is, all teachers had to find a way of introducing ideological elements in their classes, of brainwashing the students and turn them into the 'new persons' the party expected them to become. As far as English language teaching was concerned, textbooks were written with this objective in mind. For example, in the 80s there were practically no texts about British or American cultures and civilisations in school textbooks, but there were texts about the Romanian national anthem, the co-operativisation of agriculture and the glorious life of the worker in the Romanian state factories. Teaching methods were also imposed centrally and required memorising, reproduction of knowledge and drilling.

It is not surprising that foreign languages (both teaching and testing them) were considered unimportant. The Bacalaureat (which was in existence at that time, too) was the only high-stakes, nationwide, external exam. No paper in a foreign language was ever included in the Bacalaureat (except for the 1950s, when there was a compulsory Russian paper). Only in the post-1989 period has a compulsory foreign language oral test component been introduced, while an elective foreign language written test has been (and still is) also available. English has been the most popular choice of a foreign language in the Bacalaureat to-date.

Although the structure of each paper was issued centrally, the marking was done locally by teachers from the same school. The marking was highly subjective: no marking schemes were imposed or even produced locally, no reliability procedures were adopted, and no monitoring procedures were employed. It is no wonder, under the circumstances, that some of the stakeholders involved (mainly the students and their parents) were highly suspicious of the quality and fairness of the marking. A system of unethical interventions developed: parents frantically sought relatives and 'connections' who would facilitate these interventions, a practice which would eventually ensure high marks for their children. Unfortunately, this practice was so deeply rooted in the stakeholders' minds that it has become one of the most dangerous legacies of the totalitarian system, still affecting the perceptions of Romanians nowadays.

Romania was the only country in the region where the communist regime fell as a consequence of a bloody uprising. The transition from one of the most oppressive, closed totalitarian regimes in the region to an ideological and economic opposite: pluralism and market economy was therefore more dramatic and controversial than elsewhere. People expected immediate changes with immediate positive results, expectations raised by the bloodshed that had taken place.

This dramatic transition was bound to influence education. Some of the most evident and detested effects of the communist regime were almost immediately eliminated: indoctrination, ideology and excessive manifestation of power and control over persons and institutions. Nevertheless, this elimination was sometimes done only at the surface level. The educational system today still suffers from its totalitarian legacy.

An interesting book, entitled *Romanian Education Today: A Diagnostic Study*, edited by Miroiu (1998), paints a grim picture of the system of education in Romania in the late 90s, insisting on the need for change and reform. When discussing the Bacalaureat, Miroiu (1998:55; author's translation) argues:

> 'The Baccalaureate the students have to sit when leaving school is "baroque": extremely homogeneous, with papers focused on memorising and reproduction. The verification of knowledge seen as content represents an end to this examination, the contextual use of this knowledge being completely disregarded (in 1999 this aberration will be reinforced by the publication of the actual tasks together with the key, six months before the Baccalaureate, and in this way any chance to resort to strategies other than rote-learning is entirely ruled out'.

What all this leads to is that at the end of these two exams a serious and devoted student is exhausted. According to Miroiu though 'the blessed forgetting of the majority of the information so painstakingly memorised takes place only a few weeks later' (1998:66; our translation). This requirement of faithfully memorising and reproducing knowledge has encouraged a vast industry of private tutorials which, 'if things do not change (especially the system of evaluation and curricula), will focus on the Baccalaureate' (*ibid*; our translation).

The case of the Bacalaureat

As we will briefly show, the Romanian Bacalaureat has constantly suffered, since 1989 to date, from yo-yo effects. For a brief period of time, in the early 1990s, a powerful de-centralising force swept over almost all domains of Romanian social life, a tendency springing naturally from a newfound freedom. Additionally, more and more universities started to include Bacalaureat grades in their selection procedures. In the mid 90s there were already a number of universities which selected their students solely on their Bacalaureat grades.

Under the circumstances, the future status and currency of the Bacalaureat was unclear. The format continued to be dictated centrally, but the content of each paper was decided by the teachers from each secondary school and the marking continued to be done by the same teachers. Some claimed that each secondary school should have the autonomy to decide the format and content of the Bacalaureat. Others said that this would trigger major problems, such as, for instance, lack of similar standards, and argued that there was need for a unique, highly-standardised Bacalaureat (both in structure and in marking).

In spite of these contradicting views, there was agreement in one respect: the existing Bacalaureat was neither relevant nor trustworthy. In other words, worries were expressed about the validity and reliability of this important exam. These worries prompted the Ministry and the Parliament to take action. In 1995 the new Law of Education [*Legea Învăţământului*, Art. 26] legalised the structure of the Bacalaureat: what papers were to be sat and whether they would be in an oral or a written form (for example, the Romanian language was the only subject tested both in a written paper and in an oral one, while the other subjects were either taken in writing or orally). Laws

in Romania cannot be implemented without an accompanying 'methodology of application'. Thus, Article 26 stipulates that this 'methodology' was to be designed and issued yearly by the Ministry of Education and would refer to the syllabuses, administration and marking for each of the papers. This methodology was supposed to be made public before the beginning of each academic year. For this reason, each and every year the methodology could, at least in theory, contain 'new things' that the stakeholders should be informed about in sufficient time.

The National Assessment and Evaluation Service (NAES), a specialised institution funded by the World Bank, was established in 1996. Its main role was to develop reformed standardised external examinations by July 2000. Thus it was hoped that starting with July 2000 at the latest, in Romania the two external high-stakes examinations would be standardised, valid, reliable and EU-compatible.

Unfortunately, this first implementation was postponed each year until July 2003, when it was finally introduced. The major change was that all the items in all the papers were to be multiple-choice based. Moreover, it had finally come to what Miroiu (1998) had called 'the 1999 aberration': the prior publication of all the tasks and items of the Bacalaureat. The actual papers were to be constructed by selecting from those items. A great scandal followed the 2003 Bacalaureat, teachers and students took to the streets and complained that (particularly when it came to the English papers) the tasks and their answer keys were full of mistakes and unclear and that the marks obtained were unreliable. Television news reported that students with good results in internationally recognised language tests (such as Cambridge ESOL's Certificate in Advanced English or TOEFL) got very poor grades in the Bacalaureat. Consequently, in 2004 this 'new Bacalaureat' was dropped and the old one was reinstated.

In 2006 the yo-yo effect struck once again as it was decided that it was better to have all the tasks and items for all papers posted on the internet because in this way fraud (at least, as far as the content of the papers was concerned) would be prevented.

In March 2007 the NAES was dissolved and a new body was created: Centrul Naţional pentru Curriculum şi Evaluare în Învăţământul Preuniversitar [the National Centre for Curriculum and Evaluation in Secondary Education]. The aims of this body, as presented on the official site of the Ministry of Education, Research and Youth (www.edu.ro), are to deal with:

- curricular design
- coordination and organisation of the national examinations
- coordination of textbook evaluation
- selection, preparation and stimulation of valuable young people.

On the same site one can find information concerning the July 2008 Bacalaureat, namely:

- the syllabus (skills, content, types of source texts, types of target texts, themes and topics, functions and language structures)
- the formats (strands, oral and written papers)
- methodologies for administration
- the tasks and items for each paper.

As far as English is concerned, an oral paper in a foreign language is compulsory and qualifies successful students for university entry (failing would automatically trigger disqualification). Statistically, the largest number of Bacalaureat candidates in Romania choose English as their compulsory foreign language oral paper. However, a number of candidates, especially those who have studied English intensively, do choose another foreign language for their compulsory foreign language paper, and choose English for the optional written paper. On the day of the examination the computer draws a number of items for the construction of actual papers.

Progress has been made though and one cannot help noticing that the information available (at least online) for the Bacalaureat has improved both in quantity and in quality, over the years. Besides the specifications for the 2008 version, one can also find reports and statistics on the 2004–6 versions, as well as past papers and answer keys from the 2006 version. Nevertheless there are still areas in need of development and research. For example, one problem in desperate need of attention is that of task design and item writing, with particular reference to the validity of the content and to the reliability of the marking. It is not clear how the items are produced, by whom or what quality assurance strategies are employed. As far as the marking is concerned, if tasks are not marked objectively, the mark scales used are unclear, if given, and the markers are not standardised.

In brief, after looking at the nature and status of the Bacalaureat, we believe that educators in Romania, especially those in positions of power still do not know what the holder of such a qualification is supposed to do with the knowledge needed to pass it, when it comes to further studies or career. More sadly though, Romanian educators seem not to care about the stakeholders' opinions and attitudes concerning the Bacalaureat, since no nationwide survey has ever been conducted.

CAE in Romania

The CAE is by far the most widely taken external examination in Romania and the number of people sitting it grows every year. Just to give an example, in the June 2008 session (which is the most prolific session in Romania), around 3,500 candidates sat this exam. Roughly speaking 3,000 chose one of the three open-centres existing in Romania (Bucureşti, Cluj, and Iaşi) while about 500 took the exams locally in their schools. In comparison, there were only about 600 FCE candidates in the same session.[2]

There seem to be two reasons why the CAE is the most popular non-Romanian examination. The first, we believe, is a British Council and Cambridge ESOL project launched in the late 90s. This project offered opportunities and facilities for English bilingual secondary schools to organise CAE examinations for their Year 11 students. The second, we believe, is related to the decision of some prestigious state

2. To complete the picture, around 220,000 candidates sat the Bacalaureat, of whom over 80% chose at least one English paper.

universities in Romania to replace the Bacalaureat results with the CAE results in their admission schemes, when it comes to English. To our knowledge there are at least two such universities: University of Cluj and University of Timisoara.

The nature and current status of the Bacalaureat as well as the increasing popularity and currency of the CAE prompted us to conduct a small-scale survey concerning the attitudes of teachers and students towards these two examinations. The study and the most important findings are presented here.

The study

The study presented in this article is a based on a survey which was conducted in *Universitatea de Vest*, one of the most prestigious universities in Romania, at the beginning of the 2nd semester of the academic year 2007/8, in mid-February 2008. A questionnaire was administered to all 25 teachers in the Department of English and 15 questionnaires were returned. As far as the student population is concerned, the questionnaire was administered, after a course, to all the first year students in Applied Modern Languages who were present. There were 69 students present (from the 95 enrolled in this programme) and 43 agreed to complete and return the questionnaires there and then. For validation purposes, we analysed the responses separately and then we compared our findings.

Questions and findings

We will now present the most relevant questions and the answers obtained. A small number of questions could only be found in the questionnaires addressed to the teachers, others were common to both the teachers and the students.

Questions for teachers only

In this section we present the questions for teachers only and their responses. The first two questions referred to the teachers' familiarity to both exams being studied. They were asked:

1. How familiar are you with Cambridge ESOL's CAE?
2. How familiar are you with the English component of the Bacalaureat?

See Table 1 for the results.

Table 1: The teachers' familiarity with the English component of the Bacalaureat and CAE

Examination	Very	Fairly	Vaguely	Not at all
Bacalaureat	2	5	7	1
CAE	9	5	1	0

It is notable that the teachers consider that they are more familiar with the CAE than the Bacalaureat even though the Bacalureat is a national compulsory examination for admittance to tertiary education. The responses to the next two questions seem to confirm this strange attitude of the teachers when it comes to preparing the students for these two examinations (see Table 2):

3. Have you ever prepared students for the English papers in the *Bacalureat*?
4. Have you ever prepared students for a Cambridge examination?

Table 2: Teachers' preparation of students for the Bacalaureat and CAE

Examination	Yes	No
Bacalaureat	9	6
CAE	13	2

The teachers who responded had prepared students for the CAE in a much larger proportion than for the Bacalaureat. This apparent disinterest in preparing students for the Bacalaureat will be confirmed and partly explained by the findings presented below.

Questions for both teachers and students

In this section we present the most important findings concerning the attitudes of both teachers and students surveyed. Firstly, in response to the following:

1. Have you ever sat a Cambridge exam?
2. If yes, which one/s? Please give your reason(s) why you decided to sit it/them.

Table 3: Upper Main Suite Examinations sat

	FCE	CAE	CPE	Total no. of people with a Cambridge ESOL exam
Teachers (15)	0	0	7	7
Students (43)	2	7	0	9

Table 3 shows that a greater proportion of the teachers surveyed have sat a Cambridge ESOL exam than the students, almost half of the teachers sat one as opposed to 20% of students. The most popular reasons for sitting a Cambridge ESOL examination are:

- For teachers: to become an oral examiner (3 cases), international certification (1), studies (1), career (1)
- For students: international certification (6), check level (4), career (2), studies (1).

3. Did you take the English component of the Bacalaureat?

All the students responded that they chose an English paper in the Bacalaureat whereas 10 out of fifteen teachers responded positively. The five teachers who said *no* are probably graduates of the pre-1989 period when a foreign language component was neither compulsory nor an elective.

We consider the next questions as being the most relevant when it comes to the teachers' and students' attitudes towards the two examinations under scrutiny:

4. Do you think Cambridge exams are: relevant/useful/trustworthy? Please explain your choices.
5. Do you think the English component of the Bacalaureat is: relevant/useful/trustworthy? Please explain your choices.

As Table 4 clearly shows, both the teachers and the students have a more favourable attitude towards the CAE

Table 4: Features of Cambridge ESOL's CAE and the English component of the Bacalaureat

	Relevant		Useful		Trustworthy	
	CAE	**BAC**	**CAE**	**BAC**	**CAE**	**BAC**
Teachers (15)	Yes: **14**	Yes: **5**	Yes: **13**	Yes: **14**	Yes: **14**	Yes: **3**
	No: **0**	No: **10**	No: **0**	No: **1**	No: **0**	No: **11**
	I don't know: **1**	I don't know: **0**	I don't know: **1**	I don't know: **0**	I don't know: **0**	I don't know: **0**
	Not completed: **0**	Not completed: **0**	Not completed: **1**	Not completed: **0**	Not completed: **1**	Not completed: **2**
Students (43)	Yes: **24**	Yes: **18**	Yes: **41**	Yes: **40**	Yes: **25**	Yes: **20**
	No: **2**	No: **23**	No: **0**	No: **1**	No: **1**	No: **21**
	I don't know: **8**	I don't know: **0**	I don't know: **2**	I don't know: **2**	I don't know: **6**	I don't know: **2**
	Not completed: **9**	Not completed: **2**	Not completed: **0**	Not completed: **0**	Not completed: **11**	Not completed: **0**

than to the English papers in the Bacalaureat, especially when it comes to relevance and trustworthiness. However, the overwhelming majority of the teachers and students consider the Bacalaureat useful. An explanation for this apparent paradox can be found in the reasons given (see Table 5). It can be observed that one reason mentioned is the powerful role of the Bacalaureat. This is one good example which definitely supports Spolsky's claim (1997:242) that 'since their invention, tests and examinations have been a means of control and power'.

It is also interesting to observe that most of the comments from the teachers and students were of a positive nature when it comes to the CAE and of a negative nature when it comes to the Bacalaureat.

The last question from our study presented here refers to the tertiary education selection role that these examinations should have in the teachers and students' views:

6. In your view, which of the following exams should be used as selection instruments for becoming a student in an English department at the university? Cambridge ESOL's CAE/The English component of the Bacalaureat/Both of the above/None of the above/I don't know/Other(s)

See Table 6 for a summary of responses.

Table 5: A checklist of the reasons given

	Yes/ + comments		No/ – comments	
	CAE	**BAC**	**CAE**	**BAC**
Relevant	• Cambridge ESOL's professionalism • communicative nature • it tests all skills • validity • variety of task types		• term of validity too long	• based on rote learning and reproduction • irrelevant content • not all skills tested • not sure what it assesses • old fashioned • too easy
Useful	• for career • for immigration • for language development • for university studies • improves self-esteem	• a selection factor for tertiary education		• irrelevant for real life needs
Trustworthy	• computer-based marksheet correction • fairness • papers marked in Cambridge • reduced subjectivity • security	• no reason to tamper with the results		• cheating is possible • marks too high • unclear instructions • unfair practices still widespread in Romania

Table 6: Selection instruments for admission to university

	Only the CAE	Only the Bacalaureat	Only an entrance exam	Both Cambridge ESOL and the Bacalaureat	Both a Cambridge exam and an entrance exam	All three	I don't know
Teachers (15)	6	–	5	2	2	3	–
Students (43)	11	6	2	14	–	5	6

Once again the teachers and the students seem to have more positive attitudes towards the CAE than the Bacalaureat. None of the teachers and only six students consider that only the Bacalaureat should be employed as a selection instrument for tertiary education, while a much larger number would include the CAE singly, or in combination, in the selection process.

Conclusion

As shown in the previous sections, the English papers in the Bacalaureat (the Romanian school-leaving examination) are not viewed positively by the respondents (both students and teachers). They are not seen as valid and reliable. Their attitudes are much more favourable, though, when it comes to Cambridge ESOL's CAE exam. The nature, status and history of the two examinations, presented in the Context section of this article, might constitute an explanation for the respondents' feelings.

To conclude, we would like to present a quote from a focus group we conducted and which we are currently analysing qualitatively. The quote summarises the findings of the survey presented in this article and offers one possible interpretation of these findings.

> 'Honestly, I'm sure that the English exam in the Romanian Bacalaureat isn't for advanced level in English and therefore can't be as difficult and demanding as the CAE but the marking criteria are different and I don't mean that in a positive way…
> I trust the Cambridge exam more and I know I can be sure that the result reflects my real level of English.'

References and further reading

Beresford-Hill, P (1994) Educating Homo Sovieticus, *Education Today*, 44/2, 50–54.

Fullan, M G (1995) *Understanding Educational Change*, London: Routledge.

Legea nr. 84 din 24 iulie 1995 – Legea Invătământului din România, retrieved from www.cdep.ro/pls/legis/legis_pck.htp_act?ida=6696

Matei, S G (2002) *Student Teachers as Researchers: An Inquiry-Oriented Approach to Initial Teacher Education*, unpublished PhD thesis, Exeter University.

Miroiu, A (Ed.)(1998) *Invăţământul Românesc Azi: Studiu de Diagnoză*, Iaşi: Polirom.

Shohamy, E (1997) Testing methods, testing consequences: Are they ethical? Are they fair?, *Language Testing,* 14/3, 340–349.

Spolsky, B (1997) The ethics of gatekeeping tests: what have we learned in a hundred years, *Language Testing*, 14/3: 243–247.

The revised IELTS Pronunciation scale

SACHA DEVELLE RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

Introduction

A series of articles in *Research Notes* (February 2001, July 2001, November 2001, November 2006) reported the development, validation and implementation of the IELTS Speaking Test Revision Project (1998–2001). More recent research (Brown 2006, Brown and Taylor 2006) highlighted the need for clearer specification on the Pronunciation scale. Further revision arose as a consequence of this research and the continuing evolutionary process reflecting the test's context of use.

The present article describes a two-phase study using a mixed method (quantitative and qualitative) approach to examine raters' marking behaviour and perceptions of the newly revised IELTS Pronunciation scale.

Background to the IELTS Speaking scale

The IELTS Speaking test underwent a major review at the end of the 1990s based on the findings from a number of IELTS funded research studies investigating various aspects of the Speaking test (Taylor 2007). The IELTS Speaking Test Revision Project (1998–2001) focused on two key areas from those findings: assessment reliability and test conduct. Two major changes were subsequently introduced to the Speaking component. Firstly, it provided a new assessment procedure that shifted from a conversational interview style to a more structured format; an assessment shift driven by research examining interview discourse (Lazaraton 1996a, 1996b, Brown and Hill 1998). The second major change introduced 4 rating subscales that included a separate Pronunciation criterion for the first time. The revised Speaking test was introduced in 2001. IELTS Research Reports 6 (Brown 2006) and Studies in Language Testing 19 (Taylor and Falvey 2007) document the IELTS Speaking studies that contributed to the revision process. The existence of a separate Pronunciation scale has allowed us to focus on the study of Pronunciation in IELTS Speaking.

The evolutionary process of the IELTS Speaking test was again reflected by the introduction of half band reporting for the Speaking (and Writing) module in July 2007 that allowed for a more sensitive reporting of performance judgments on each of the assessment criteria.

Revising the IELTS Pronunciation scale

More recent research carried out on the IELTS Speaking test has shown that the 4-point scale (using the even-numbered Bands 2, 4, 6, 8) designed for Pronunciation needed clearer specification for key performance features, particularly when distinguishing critical boundaries that exist above and below Band 6. A large-scale survey commissioned by Cambridge ESOL in 2005 investigated examiners' views and experiences of the revised IELTS Speaking Test. Brown and Taylor (2006) discussed some of those findings in *Research Notes* 26. A general finding

that emerged was that examiners felt the pronunciation descriptors did not discriminate sufficiently clearly between different levels of proficiency and that this was due to the constraint of having only the four Pronunciation bands (2, 4, 6, 8). Research funded by the IELTS partners (Brown 2006) also highlighted a common perception amongst examiners, namely that the lack of a full 9-point scale for Pronunciation (i.e. the absence of a Band 5 or 7) could lead to Band 6 being set as a default level.

The main purpose of the IELTS Pronunciation scale revision, therefore, was to place the rating of pronunciation on a full 9-point scale in line with the other three analytical criteria (Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy). The extension of the scale would then allow for a clearer specification on key performance pronunciation features. Table 1 compares features of the Pronunciation scale for 2001 and 2007.

Table 1: A comparison of features for the IELTS Pronunciation Scale

Pronunciation Scale 2001	Revised Pronunciation Scale 2007
Whole band reporting 4 Point (even) Scale	Half band reporting 9 Point (even/odd) Scale

A mixed method approach

A two-phased study incorporating a mixed method (qualitative and quantitative) approach examined raters' marking behaviour and perceptions of the newly revised scale. The wording for the new descriptors was of crucial importance and as such highlights the importance of the qualitative methodology incorporated into the design. Qualitative responses to the feedback questionnaire were placed into running themes and examined by narrative threads. Candidate score data was analysed using Multi-Faceted Rasch measurement (FACETS). The following research questions were formed for this study:

- What common themes run through examiner feedback responses when using the 9-point revised scale?
- Do examiners make full use of the extended scale?
- What are the levels of agreement, consistency and harshness/leniency in examiners' use of the revised scale?

Some of the findings from the qualitative and quantitative analyses are reported below.

Preliminary phase (April–June 2007)

A Pronunciation Scale Working Group was established in April 2007 that involved IELTS Principal and Assistant Principal Examiners and internal Cambridge ESOL staff with extensive experience in the Speaking component of IELTS and Main Suite exams. The wording for the first draft of additional band descriptors at Bands 1, 3, 5, 7, 9 was produced. The first draft was then presented to a group of IELTS examiners who participated in a rating pre-trial phase. Their ratings and commentary contributed to the production of the Draft 2 descriptors.

Phase 1 (July–August 2007)

Twenty four experienced IELTS examiners chosen from a range of geographical locations participated in Phase 1. Examiners were asked to rate 15 IELTS Speaking performances. Candidates included 7 males and 8 females. They came from a range of L1 backgrounds and proficiency levels and included Speaking performances ranging from Band 3 to Band 9 on the IELTS scale. Examiners were instructed to rate the 15 performances using the new 9-point assessment scale (Draft 2) and to complete a feedback questionnaire.

Feedback questionnaire

Examiners rated 4 questions on a 5-band Likert scale (Disagree=1, Agree=5) and provided commentary on 5 open-ended questions. Table 2 shows the raw score (n=24) and percentage responses to the Likert scale questions.

Responses to the first question show that 16 out of 24 raters circled 4 or 5 as their preferred choice, suggesting that these raters found the revised pronunciation scale easy to apply. Six raters (25%) chose the middle of the scale (3), implying that they found the scale neither easy nor difficult. Nineteen (79%) of the raters felt confident differentiating over the 9 Bands. This was a very positive response as it suggests the wording for each descriptor allows raters to distinguish between bands. Raters were also confident in distinguishing between Bands 8 and 9 as shown by 19 out of 24 (79%) choosing either 4 or 5 on the scale. Sixteen out of 24 raters (67%) circled 4 or 5 for confidence in applying a pronunciation rating. However, a closer look at those ratings showed that a third of the raters also chose 2 or 3. This last finding suggests a lack of confidence when using the revised scale. The following qualitative feedback from open-ended questions provides some insight into why this could be the case.

Written responses to the 5 open-ended questions were divided into running themes and analysed for specific comments related to ease and/or difficulty in applying the new descriptors. Overall qualitative feedback was consistently more positive than negative. Question 8 asked raters to put forward their own comments on the

Table 2: Phase 1 responses to Likert scale questions (n=24)

Statement	Pronunciation Scale				
	Selecting				
	1	2	3	4	5
The revised scale was easy to apply	0 (0%)	2 (8.3%)	6 (25%)	14 (58.4%)	2 (8.3%)
It is possible to differentiate over the 9 Bands	0 (0%)	1 (4.2%)	4 (16.7%)	10 (41.6%)	9 (37.5%)
It is possible to distinguish between Band 8 and 9	0 (0%)	1 (4.2%)	4 (16.7%)	6 (25%)	13 (54.1%)
I am confident in my Pronunciation rating	0 (0%)	2 (8.3%)	6 (25%)	13 (54.2%)	3 (12.5%)

revised scale. A representative sample of responses included:

> It will be good to have a 9 point scale.
>
> I actually found it much easier to distinguish the higher and lower levels with the new descriptors.
>
> I am very much in favour of these new additions.
>
> A very positive change.
>
> The new band descriptors are much better and take the examiners' attention away from general intelligibility and attempt to focus attention on phonological features.

However, there were common narrative threads that highlighted the need for further revision to the existing descriptors for Bands 4 and 6. The following comments are a representative sample:

> The descriptors for Bands 4 and 6 contain many negative features and few positive features, making it difficult to apply ratings to Bands 3 and 5.
>
> The descriptors for Bands 4 and 6 contain a majority of negative features, so it is perhaps not appropriate to say 'displays some but not all of the features' when describing a lower band.
>
> There was difficulty applying Band 3 as there are so few positive features at Band 4.

These last comments perhaps highlight the reason for why a third of the respondents (displayed in Table 2) circled 2 or 3 for confidence in applying Pronunciation ratings.

Facets analysis

The Multi-Faceted Rasch measurement provides harshness and leniency estimates and levels of consistency for raters. The issue of what is an acceptable range of examiner severity is quite complex and there are no universally accepted rules. Van Moere (2006) provides a range of -1.00 and +1.00 logits as a useful cut-off point. Applying these standards, the majority of raters fell within the acceptable range of rater severity. There was some variation amongst raters as a group with differences shown by the most lenient (-1.42) and harshest (1.47) logit scores.

Rater consistency levels were compared with Wright and Linacre's (1994) suggested range of 0.6–1.4. Two of the 24 raters fell just outside (1.72 and 1.73 respectively). However these scores were not seriously misfitting and show that as a group raters performed in a consistent manner in their use of the overall scale. The point biserial correlation on the Pronunciation criterion, based on the agreement between one rater and all others, ranged from .87 to .96 which is encouragingly high. Finally, Speaking performance ratings showed a good spread of scores (Bands 3–8).

To briefly sum up, the findings from the quantitative analysis showed the majority of raters performed consistently, fell within the accepted parameters of rater severity and used the full length of the scale. The qualitative findings were also predominantly positive. However, the need to revisit the wording for Bands 4 and 6 emerged as an important issue to be re-addressed and was the impetus behind Phase 2 reported below.

Phase 2 (September–October 2007)

Further revision to the wording of Descriptors 4 and 6 was carried out by the IELTS Pronunciation Working Group described in Phase 1. The production of Draft 3 contributed to Phase 2. A different set of 10 experienced examiners participated in the second phase. Twenty taped IELTS Speaking performances were chosen from a range of L1 backgrounds and proficiency levels. Each rater received Draft 3 of the revised Pronunciation scale, the Speaking performances and the feedback questionnaire.

Feedback questionnaire

Table 3 displays the raw score and percentage results from the 5-band Likert Scale (Disagree=1, Agree=5).

Nine out of ten raters found the revised scale easy to apply. All 10 raters (100%) found it possible to differentiate over the 9 Bands, suggesting that earlier concerns in applying a rating for Bands 3 and 5 (due to the original wording of Bands 4 and 6) were not an issue for raters in Phase 2.

Table 3: Phase 2 responses to the Likert scale questions (n=10)

Statement	Pronunciation Scale				
	Selecting				
	1	2	3	4	5
The revised scale was easy to apply	0 (0%)	0 (0%)	1 (10%)	7 (70%)	2 (20%)
It is possible to differentiate over the 9 Bands	0 (0%)	0 (0%)	0 (0%)	5 (50%)	5 (50%)
It is possible to distinguish between Band 8 and 9	0 (0%)	0 (0%)	2 (20%)	7 (70%)	1 (10%)
I am confident in my Pronunciation rating	0 (0%)	0 (0%)	2 (20%)	8 (80%)	0 (0%)

Eight out of 10 raters circled 4 or 5 on the Likert scale when asked if it was possible to distinguish between Bands 8 and 9. There were 8 out of 10 raters that circled 4 in terms of confidence in applying Pronunciation ratings. This result suggests that the re-wording for Descriptors 4 and 6 allowed for clear distinctions to be made on the overall scale. Written responses to the 5 open-ended questions were again divided into running themes. There was a particular focus on identifying any comments on Bands 4 and 6. The absence of any reference to those descriptors was very encouraging as their rewording was the primary focus of Phase 2. Overall, the analysis of responses showed that the revised IELTS Pronunciation scale met with a high level of approval from raters. The most frequent comments included:

> Very useful to have more detail on prosodic features.
>
> It's not the 9 bands but the more detailed descriptors which are most helpful.
>
> Initially I felt that the new descriptors for Band 5 and 7 were vague, but as I used them I realised that they work really well and are exactly what examiners need to be more specific.

Facets analysis

Results from the FACETS analysis showed that examiners made full use of the extended Pronunciation scale. Speaking performance ratings showed a spread of scores from Band 4 to Band 9. In terms of consistency, as a group examiners rated very uniformly (.43–1.47) on the Pronunciation criterion. There was very little variation for harshness/leniency amongst the group with 9 of 10 raters falling within Van Moere's (2006) acceptable range of -1.00 and +1.00 (-.62 to +.67 logits). The point biserial correlation for Pronunciation, based on the agreement between one rater and all others, ranged from .85 to .95 which is encouragingly high.

Discussion

The present article has presented key findings from Phase 1 and Phase 2 of the revised Pronunciation scale project carried out between April–October 2007. We now return to the original research questions. In terms of common themes from the qualitative questionnaire, the iterative feedback from raters during Phase 1 and Phase 2 guided final decisions on wording for descriptors. Phase 2 findings showed that raters generally felt confident using the new scale and found no difficulty in distinguishing between bands. The re-wording of descriptors (Draft 3) presented during Phase 2 also produced an interesting finding in terms of rater assessment. For example, in Phase 1 the open-ended Question 4 (*When assessing Pronunciation with the new scale what do you generally find yourself paying most attention to?*) resulted in a variety of responses from raters. This response was in contrast to the same question in Phase 2 that produced a unified focus on specific pronunciation features (e.g., stress, intonation and rhythm, chunking, pronunciation of individual words). It also seems that the rewording of Draft 3 descriptors helped raters to anchor their assessment strategies.

Our second research question investigated how examiners made use of the extended pronunciation scale. The results from both phases were very positive with raters employing the full range of the scale. The final research question focused on levels of consistency, agreement and harshness/leniency findings for raters. The results from Phase 1 and 2 showed that the majority of raters were consistent in using the revised Pronunciation scale. Levels of agreement between raters were encouragingly high and the majority of raters fell within the acceptable boundaries in terms of rater severity.

Conclusion

The IELTS Speaking test has passed through a series of evolutionary stages, and with each new dimension the test qualities pertaining to Cambridge ESOL's VRIP framework embracing Validity, Reliability, Impact and Practicality continue to be evaluated. The revised Pronunciation scale became fully operational in August 2008 and monitoring of Speaking test performance data, including the functioning of the Pronunciation scale, will continue to form part of the ongoing IELTS research and validation agenda.

References and further reading

Anderson-Hsieh, J (1995) Pronunciation factors affecting intelligibility in speakers of English as a foreign language, *Speak Out*, 18, 17–19.

Brown, A (2006) An examination of the rating process in the revised IELTS Speaking Test, in *IELTS Research Reports,* volume 6, IELTS Australia/British Council.

Brown, A and Hill, K (1998) Interviewer style and candidate performance in the IELTS oral interview, in *IELTS Research Reports*, volume 1, Sydney: ELICOS, 1–19.

Brown, A and Taylor, L (2006) A worldwide survey of examiners' views and experience of the revised IELTS Speaking test, *Research Notes,* 26, 14–18.

Celce-Murcia, M, Brinton, D and Goodwin, J (1996) *Teaching pronunciation: A reference for teachers of English to speakers of other languages*, Cambridge: Cambridge University Press.

Derwing, T M and Munro, M J (1997) Accent, intelligibility and comprehensibility: Evidence from four L1s, *Studies in Second Language Acquisition*, 20, 1–16.

DeVelle, S and Galaczi, E (2007a) *Revising the IELTS Pronunciation scale 2007: Phase 1*, Cambridge ESOL internal report.

— (2007b) *Revising the IELTS Pronunciation scale 2007: Phase 2*, Cambridge ESOL internal report.

Gustafson, J E (1980) Testing and obtaining fit of data to the Rasch model, *British journal of mathematical and statistical psychology*, 33, 220.

Horne, M (Ed.) (2000) *Prosody: theory and experiment*, Dordrecht: Kluwer Academic.

Hurley, D (1992) Issues in Teaching Pragmatics, Prosody, and Non-Verbal Communication, *Applied Linguistics*, 13/3, 259–280.

Lazaraton, A (1996a) A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE), in Milanovic, M and Saville, N (Eds) *Performance Testing, Cognition an Assessment: Selected papers from the 15th Language Testing Research Colloquium* (Studies in Language Testing volume 3), Cambridge: UCLES/Cambridge University Press, 18–33.

— (1996b) Interlocuter support in oral proficiency interviews: The case of CASE, *Language Testing,* 13, 151–172.

— (2002) *A qualitative approach to the validation of oral speaking tests*, (Studies in Language Testing volume 14), Cambridge: UCLES/Cambridge University Press.

Morley, J (1991) The pronunciation component of teaching English to speakers of other languages, *TESOL Quarterly*, 25, 481–520.

Taylor, L (2001a) Revising the IELTS Speaking test: developments in test format and task design, *Research Notes*, 5, 2–5.

— (2001b) Revising the IELTS Speaking test: retraining IELTS examiners worldwide, *Research Notes*, 6, 9–11.

— (2007) The impact of the joint-funded research studies on the IELTS Speaking module, in Taylor, L and Falvey, P (Eds) *IELTS Collected Papers: Research in Speaking and Writing Asessment* (Studies in Language Testing volume 19), Cambridge: UCLES/Cambridge University Press, 185–194.

Taylor, L and Jones, N (2001) Revising the IELTS Speaking test, *Research Notes*, 4, 9–12.

Upshaw, J A and Turner, C E (1999) Systematic effects in the rating of second language speaking ability: Test method and learner discourse, *Language Testing*, 16, 84–111.

Van Moere, A (2006) Validity evidence in a universal group oral test, *Language Testing*, 23/4, 411–440.

Wright, B D and Linacre, J M (1994) Reasonable mean-square fit values, *Rasch Measurement Transactions*, 8/3, 370.

Recent publications of interest

IELTS Research Reports, Volumes 6, 7 and 8

Between 1998 and 2003, IELTS Australia published five volumes of IELTS Research Reports covering projects completed under the joint-funded research program. In 2006 Volume 6 in the series was the first of two volumes to be produced and published jointly by IELTS Australia and the British Council. It contains seven reports of research studies focusing on the IELTS Speaking test conducted between 2002 and 2004 by applied linguists and language testers in Australia, New Zealand and the UK. A follow-up volume in 2007 – Volume 7 – brought together a set of six empirical studies, four focusing on aspects of IELTS impact and consequential validity, and two on the IELTS Writing test, specifically the nature of writing performance across different proficiency levels and under different test conditions.

This year a further five studies have been published by IELTS Australia as Volume 8 in the IELTS Research Reports series. The studies all received grant funding under Rounds 10 and 11 (2004–5) of the joint-funded program. They explore issues of test washback and impact which are themes that continue to attract attention and provoke debate among the language testing community and beyond, especially in relation to high-stakes and widely-taken international proficiency tests such as IELTS.

Volumes 6, 7 and 8 are available in both hard copy and CD-Rom versions. For more information, visit the IELTS website: www.ielts.org

Studies in Language Testing

October 2008 saw the publication of Volume 27 in the *Studies in Language Testing* series, published jointly by Cambridge ESOL and Cambridge University Press. Volume 27, edited by Lynda Taylor and Cyril J Weir, is entitled *Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference May 2005*.

This latest volume brings together a collection of edited papers based on presentations given at the 2nd ALTE Conference, held in Berlin in 2005, in order to explore the impact of multilingualism on language testing and assessment. The 20 papers consider ways of describing and comparing language qualifications to establish common levels of proficiency, balancing the need to set shared standards and ensure quality, and at the same time sustain linguistic diversity. The contributions address substantive issues in assessing language ability today, grouped according to three broad themes: Section One examines issues of *transparency and diversity* and especially the role of the CEFR; Section Two considers *quality and diversity* through the application of testing standards and codes of practice and research into examination impact; and Section Three focuses on *ethics and diversity*, and the complex relationships between linguistic identity and diversity on one hand, and immigration and citizenship policy on the other.

With its broad coverage of key issues, combining theoretical insights and practical advice, this volume will be a valuable reference for academics, policy-makers and practitioners. More information is available at: www.cambridgeesol.org/what-we-do/research/silt.html

Publications by ESOL research staff

An article by Evelina Galaczi recently appeared in the peer reviewed journal *Language Assessment Quarterly* (Issue 5/2, pp 89–119). In her paper entitled 'Peer–peer interaction in a speaking test: The case of the First Certificate in English Examination', Evelina reports a discourse-based study exploring paired test-taker discourse in the First Certificate in English speaking test. The study's primary aim was to focus on fundamental conversation management concepts, such as overall structural organisation, turn-taking, sequencing, and topic organisation found in the dyadic test-taker interaction in 30 pairs of test takers. The analysis highlighted global patterns of interaction, termed Collaborative, Parallel and Asymmetric, and salient features of interaction characteristic to each pattern. A second goal of the study was to investigate a possible relationship between the patterns of peer–peer interaction and test scores for the assessment criterion Interactive Communication. The results showed a close relationship between type of talk and test scores, thus providing validity evidence for the test scores. The study discusses implications for a more accurate understanding of the construct of conversation management underlying the FCE examination, and for the empirically-based construction of assessment criteria for speaking tests.

Szilvia Papp co-edited a recently published volume entitled *Linking up Contrastive and Learner Corpus Research* in the series "Language and Computers – Studies in Practical Linguistics", published by Rodopi. The volume's editors – Gilquin, Díez Bedmar and Papp – have brought together a collection of papers that investigate the relation between contrastive analysis and interlanguage analysis, using authentic written native and foreign/second language learner corpus data representing various languages (English, Chinese, Dutch, French, German, Greek, Italian, Polish, Spanish and Swedish). The phenomena dealt with range from syntax (adverb placement, the article system, postverbal subjects) to lexis (collocations) through discourse (information packaging, thematic choice); there is also a section on methodological issues. Two of the papers in the volume were co-authored by Szilvia – one on 'The use of the English article system by Chinese and Spanish learners', and another on 'Creativity in the use of verb + noun combinations by Chinese learners of English'. The book will be of interest to a wide range of readers, especially professionals working in foreign language teaching and assessment, as well as materials writers, and researchers in second language acquisition or contrastive linguistics.