UNIVERSITY *of* CAMBRIDGE
ESOL Examinations

Research Notes

Issue 42/November 2010

A quarterly publication reporting on research, test development and validation

Senior Editor & Editor

Dr Hanan Khalifa & Dr Ivana Vidaković

Editorial Board

Dr Nick Saville, *Director,* Research and Validation Group, Cambridge ESOL

Roger Johnson, *Director*, Assessment and Operations Group, Cambridge ESOL

Production Team

Caroline Warren, Research Support Administrator

Rachel Rudge, Production Controller

George Hammond, Design

Printed in the United Kingdom by Océ (UK) Ltd.

Research Notes

Contents

Editorial Notes

Welcome to issue 42 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

This special issue of *Research Notes* shares with the readers summaries of doctoral and Master's theses by Cambridge ESOL staff. The issue is organised according to skill area and domain of interest. It begins with Nick Saville's paper on an expanded impact model intended to provide a more effective way of understanding how language examinations impact on society. In the area of reading, Hanan Khalifa investigates the construct validity of the reading module of an EAP test battery using qualitative and quantitative research methods. Also using a mixed-method approach, Karen Ashton compares reading proficiency levels of secondary school learners of German, Japanese and Urdu, while Angela Wright examines context validity of the ICFE test of Reading. If your interests lie in the area of speaking, you may want to read Mark Elliott's paper on affective factors in oral communication, Evelina Galaczi's summary of her thesis on paired test format and Ivana Vidaković's summary on learning how to express motion in a second language and factors affecting second language acquisition. In the area of writing, we would like to introduce to you Graeme Bridges' paper on cognitive validity of IELTS, Sian Morgan's paper on qualification and certainty in L2 writing, Gad Lim's work on prompt and rater effect in assessing writing, Lucy Chambers' summary on comparability issues between paper-based and computer-based modes of assessment and Hugh Bateman's work on context and cognitive validity of a BEC Writing paper. Finally, Juliet Wilson discusses models of teaching supervision, Marylin Kies proposes a framework for assessing and comparing examinations linked to the CEFR and Muhammad Naveed Khalid investigates IRT model fit from a variety of perspectives.

We finish this issue by reporting on the conference season and events Cambridge ESOL supported. Laura Cope and Tamsin Walker report on the IACAT conference (June 2010) on computerised adaptive testing. Martin Nuttall describes the ALTE events and Lynda Taylor provides a brief on the three latest volumes in the SiLT series.

Developing a model for investigating the impact of language assessment

NICK SAVILLE RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

Introduction

This summary is based on a doctoral thesis submitted to the University of Bedfordshire (UK) in 2009. Financial support for the PhD was provided by Cambridge ESOL and it was supervised by Professor Cyril Weir.

The main research question was:

> What are the essential components of an action-oriented model of impact that would enable the providers of high-stakes language examinations to investigate the impact of their examinations within the educational contexts in which they are used?

The thesis was based on the premise that there is no comprehensive model of language test or examination impact and how it might be investigated within educational contexts by a provider of high-stakes examinations. It, therefore, addressed the development of such a model from the perspective of Cambridge ESOL as a provider of English language tests and examinations in over 150 countries.

The starting point was a discussion of examinations within educational processes generally and the role that examinations boards, such as Cambridge ESOL, play within educational systems. The historical context and assessment tradition were an important part of this discussion.

In the literature review, the effects and consequences of language tests and examinations were discussed with reference to the better known concept of washback and how impact can be defined as a broader notion operating at both micro and macro levels. This was contextualised within the assessment literature on validity theory and the application of innovation theories within educational systems.

The stance in this work reflected the author's own interests and responsibilities in developing a model of impact to guide practice within the organisation. His voice as participant, reviewer and developer of the impact model, as well as his relationships with other participants and researchers, were an important feature of this work and its methodological framework. Starting in the early 1990s a series of projects were carried out to implement an approach to impact which had begun to emerge in Cambridge ESOL at that time.

Methodologically, the research was based on a *meta-analysis* which was employed in order to describe and review three impact projects. These three projects had been carried out by researchers based in Cambridge to implement an approach to test impact which had emerged as part of the test development and validation procedures adopted by Cambridge ESOL. A differentiating feature compared with research being conducted elsewhere was the emphasis on actions and activities which would allow Cambridge ESOL to 'work for positive impact' and to avoid negative consequences for test users.

Based on the analysis, the main outcome of the thesis was an expanded model of impact designed to provide examination providers with a more effective 'theory of action'. When applied within Cambridge ESOL, this model allows anticipated impacts of the English language examinations to be monitored more effectively and leads to well-motivated improvements to the examination systems. Wider applications of the model in other assessment contexts were also suggested.

The concept of impact in language assessment

Impact is relatively new in the field of language assessment and has only fairly recently appeared in the literature as an extension of washback. Both terms were discussed in the literature review. Broadly speaking, impact is the superordinate concept covering the effects and consequences of tests and examinations throughout society, whereas washback is more limited and refers to the influence of tests and examinations in teaching and learning contexts.

The literature review covered relevant work in applied linguistics, assessment and education, mainly focusing on a 15-year period up to 2004. The notion of washback which was developed in the 1990s to take account of changing views of validity in language testing provided a useful basis for building an expanded model of impact. Much of the research in the language testing literature, however, had been small-scale projects and no systematic programme had been initiated and carried out by staff within a major examination provider.

From washback to impact

The literature review summarised the developments of washback and impact models starting with Alderson & Wall (1993) and ending with Green's (2003) washback model. See for example Cheng, Watanabe & Curtis (2004) for a useful overview.

The dimensions of the washback models which emerged in the 1990s can be summarised in the following seven points.

The test features: Surface features of the test were the main focus, for example item types and formats (e.g. multiple choice). Content validity, especially in terms of authenticity, had become an important issue. In test validation (evidence of validity) the unitary concept of validity was beginning to be adopted, in particular through the influence of Bachman (see below).

The context: There was one main context which was the focus of attention: the school and classroom (i.e. the micro context). The test-taking context was typically not separated from the school context where the teaching and learning takes place. Although some wider contextual features (macro context) were starting to be discussed, these were not yet a major focus.

The participants: The main participants were taken to be the teacher and the learners in the classroom/school context. There was a limited focus on other participants, such as materials writers, or participants from the wider context (e.g. parents).

The outcomes: Outcomes were seen as changes attributable to the introduction of the test: behaviour of participants – actions, activities, performance in the target language; views and attitudes of participants; decisions to make changes to the curriculum/syllabus and to develop new materials and methods (products).

The processes involved in bringing about the outcomes were not well understood nor well represented in the model. For example, the processes whereby the test features influenced the content and methods of the teachers were not understood. Some evidence existed to suggest that content but not the teaching methodology was affected, but when these effects occurred, how they actually came about and what factors influenced the strength of the effects was not included in the model.

The researcher: The washback researcher was typically an academic, not usually involved in the test development process as a participant, nor as a participant in the teaching/learning context itself (i.e. an outsider).

The research methods: No clear impact methodology, instrument validation procedures or validated instruments had been established, but qualitative methods were emerging in addition to survey techniques for data collection. The need to problematise washback in terms of hypotheses had been recognised.

The timeline: In the washback model, the timeline was implied but not explicitly focused on. The need for comparative data – before/after – had led to a focus on time-series designs and an appeal to insights from innovation theory. Innovation theory, in relation to Wall's (1999, 2005) work using Henrichsen's (1989) hybrid model of diffusion/implementation, suggests that each period of an educational innovation has its own antecedents, processes and consequences. The investigation of 'antecedent conditions' are Henrichsen's version of the baseline study (see also Saville 2003). The consequences, therefore, are the changes which are brought about as a result of the new processes which have been introduced.

Cheng (1997, 2005), Green (2003, 2007) and Wall (1999, 2005) looked at different aspects of washback and had begun to focus more broadly on impact issues. However, there had been no serious attempt to bring all the features of impact together within a comprehensive model which would allow the complex relationships to be examined across broader educational and societal contexts.

Locating impact research within Cambridge ESOL

A fundamental concern in the thesis was how impact-related research can be integrated into operational processes. For Cambridge ESOL, impact research needed to combine theoretical substance with practical applications and to become an integral part of the operational test development and validation processes.

In placing impact within a validation framework, the work of Bachman was influential, especially his series of seminars delivered in Cambridge in 1990–1. He was one of the first language testers to discuss impact as a 'quality' of a test and suggested that impact should be considered within the overarching concept of test usefulness (Bachman & Palmer 1996). The development of 'useful tests' involves the balancing of four qualities: validity, reliability, impact and practicality – the VRIP features as they became known in Cambridge.

In an internal working paper, Milanovic & Saville (1996) first set out ideas on an expanded concept of test impact to meet the needs of Cambridge ESOL. They addressed the question of how examinations can be developed with appropriate systems in place to monitor and evaluate their impact.

Aware of the work of Hughes (1989) and others (e.g. Bailey 1996) who used checklists of behaviours to encourage positive washback, Milanovic & Saville (1996) proposed four maxims to support working practices:

Maxim 1: PLAN
Use a rational and explicit approach to test development

Maxim 2: SUPPORT
Support stakeholders in the testing process

Maxim 3: COMMUNICATE
Provide comprehensive, useful and transparent information

Maxim 4: MONITOR and EVALUATE
Collect all relevant data and analyse as required

The statements were deliberately designed to be short and memorable, to capture the key principles and what is most relevant, and in so doing to provide a basis for decision-making and action planning.

Under Maxim 1 there was a requirement to plan effectively and for the organisation to adopt a rational and explicit model for managing the test development processes in a cyclical and iterative way. Maxim 2 focused on the requirement to provide adequate support for the stakeholders involved in the many processes associated with international examinations. Maxim 3 focused on the importance of communication and of providing useful and transparent information to the stakeholders and Maxim 4 on the requirement to collect relevant data and to carry out analyses as part of the iterative process model.

By conceptualising impact within VRIP-based validation processes, there was an explicit attempt to integrate impact research into ongoing procedures for accumulating validity evidence. The Cambridge perspective on impact was framed by these considerations and provided the starting point for the model developed in the thesis.

Locating impact within educational systems

The thesis focused broadly on how impact operates within educational systems and the literature on educational reform and management of change was particularly relevant. An understanding of how socio-political change processes work within education was also considered to be crucial (Fullan 1991).

Several concepts emerged from the literature and were explored:

- a definition of stakeholders and the roles they play in many varied contexts where language learning and assessment operate
- a view of educational systems as complex and dynamic in which planned innovations are difficult to implement successfully
- an understanding of how change can be anticipated and how change processes related to assessment systems can be successfully managed through the agency of an examination provider
- the critical importance of the evidence collected as part of the validation system and as the basis for claims about validity.

It has been suggested that educational processes take place within complex dynamic systems with interplay between many sub-systems and 'cultures' and where understanding the roles of stakeholders as participants is a critical factor (e.g. Fullan 1993, 1999, Thelen & Smith 1994, Van Geert 2007).

The thesis situated the discussion of impact within the work of researchers who focus on how change can be managed successfully within educational systems. Figure 1 illustrates macro and micro contexts within society; it shows how diversity and variation between contexts tend to increase as the focus moves from the macro context to the multiple micro contexts at the local level (i.e. schools, classes, groups, individual teachers and learners).

Understanding the nature of context within educational systems and the roles of stakeholders in those contexts are clearly important considerations for an examination board like Cambridge ESOL (see Saville 2003:60).

Using case studies as meta-data

A range of data collection and analysis techniques needs to be employed in impact-related research. These were discussed with reference to the literature on social research. Ways in which quantitative and qualitative approaches can be effectively combined in mixed-method designs were noted and the validation of instruments was illustrated.

Three case studies formed the central part of the thesis.

- **Case 1** was the survey of the impact of IELTS (the International English Language Testing System). This was the starting point for the impact model; it set out the conceptualisation of impact and described the design and validation of suitable instruments to investigate it, as applied within four Impact Projects as part of an ongoing programme of validation following the 1995 revision. This case included a description of the IELTS development and the underlying constructs, the nature of the impact data which was targeted and the necessary instrumentation to collect that data. The lessons learned were summarised in relation to the developing model and how they informed the next phase of development in Case 2.
- **Case 2** was the Italian Progetto Lingue 2000 (PL2000) Impact Study. This impact study was an application of the original model within a macro educational context and described an initial attempt at applying the approach within a state educational context, i.e. the Italian state system of education and a government reform project intended to improve standards of language education at the turn of the 21st century – the Progetto Lingue 2000. The impact of the reforms generally and the specific role of external examinations provided by Cambridge ESOL formed the basis of this case. This study provided greater

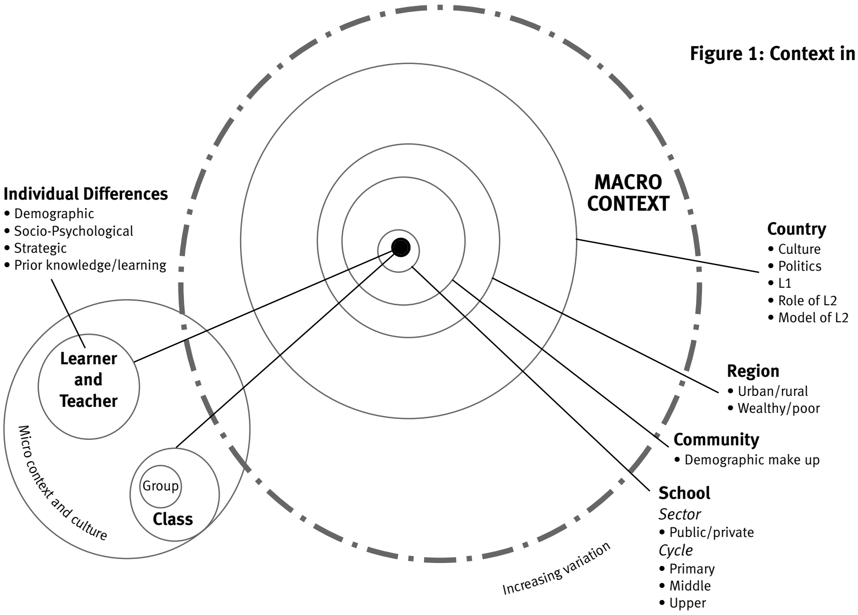


Figure 1: Context in education – A complex dynamic system

focus on the contextual variables and the roles and responsibilities of particular stakeholder groups and individuals within the educational system (see Hawkey 2006).

- **Case 3** was the Florence Learning Gains Project (FLGP). Still within Italy, this project built directly on the PL2000 case and was an extension and re-application of the model within a single school context (i.e. at the micro level). It focused on individual stakeholders in one language teaching institution, namely teachers and learners preparing for a range of English language examinations at a prestigious language school in Florence. The complex relationships between assessment and learning/teaching in a number of language classrooms, including the influence of the Cambridge examinations, were examined against the wider educational and societal milieu in Italy. The micro level of detail, as well as the longitudinal nature of the project conducted over an academic year, were particularly relevant in this case.

The analysis and discussion in each case study was broadly structured around the seven features of the washback model which had emerged by end of the 1990s, as noted above.

The revised model of impact

Insights from the three case studies were assembled into an expanded model; this meta-framework builds on Milanovic & Saville's maxims (1996), and constitutes an action-oriented approach with four inter-related dimensions (see Figure 2).

Dimension 1: re-conceptualise the place and role of impact study within the assessment enterprise, vis-à-vis societal systems generally and language education specifically.

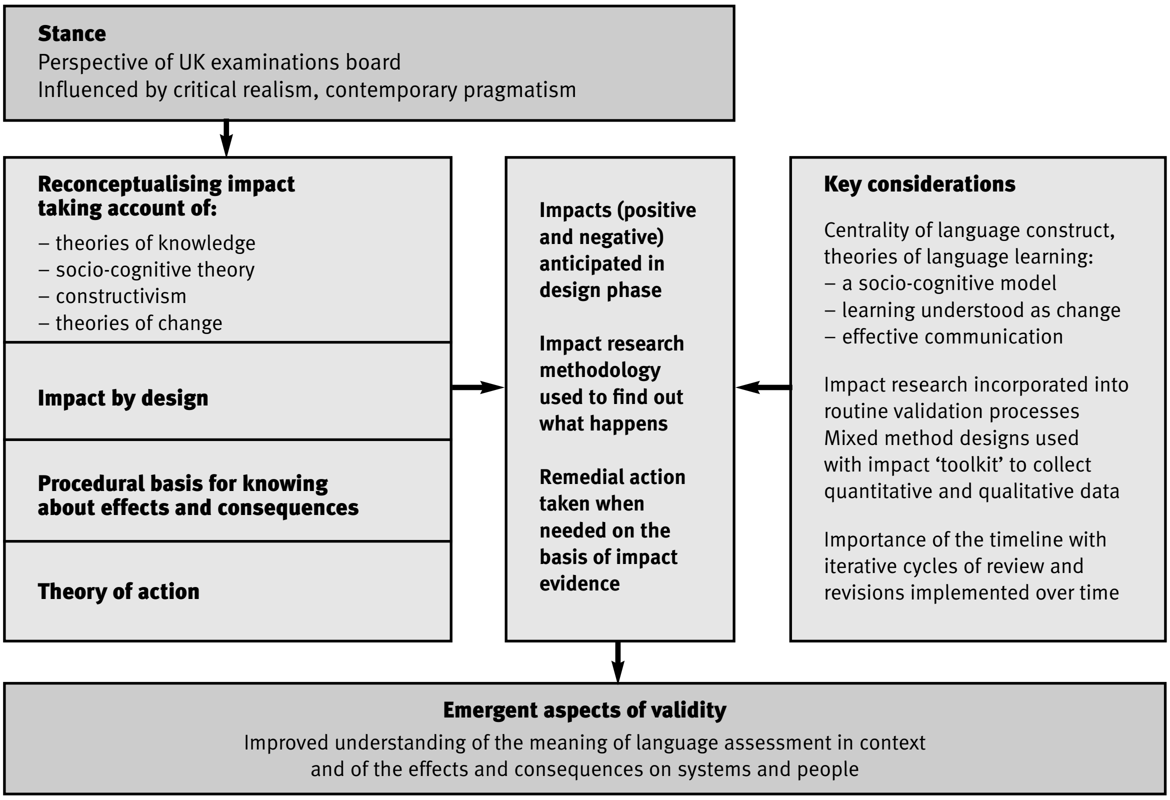
The re-conceptualisation of test impact draws on theories in the social sciences and goes beyond the work in applied linguistics and measurement. It is based on a 21st century world view and takes into account recent ontological and epistemological developments.

It extends the epistemological influences which guided Messick and his predecessors in the development of validity theory in the second half of the 20th century. Messick explicitly referred to the philosophical perspectives of Leibniz, Locke, Kant, Hegel and Singer, and to the influences of their rationalism and logical positivism on the nature of scientific enquiry in the 20th century (Messick 1989:30). In moving beyond Messick into the 21st century, the influence of post-modernism cannot be ignored, but for examinations boards and language test providers an epistemology which can provide the basis for action is required.

The ontological approach suggested draws on 'critical realism' in the social sciences (e.g. Sayer 1984, 2000) and contemporary views on pragmatism derived from the philosophy which originated with Peirce in the late 19th century. This realist stance underpins the suggested re-conceptualisation of impact and the other dimensions of the meta-framework:

a. Anticipating and managing change over time is a key aspect of impact research, noting the importance of timescales and the timeline (change over time, planned and unpredicted) with recurrent cycles (before/during/after). The recent educational literature on management

Figure 2: Revised model


Stance
Perspective of UK examinations board
Influenced by critical realism, contemporary pragmatism
Reconceptualising impact taking account of:
– theories of knowledge
– socio-cognitive theory
– constructivism
– theories of change
Impact by design
Procedural basis for knowing about effects and consequences
Theory of action
Impacts (positive and negative) anticipated in design phase
Impact research methodology used to find out what happens
Remedial action taken when needed on the basis of impact evidence
Key considerations
Centrality of language construct, theories of language learning:
– a socio-cognitive model
– learning understood as change
– effective communication
Impact research incorporated into routine validation processes
Mixed method designs used with impact 'toolkit' to collect quantitative and qualitative data
Importance of the timeline with iterative cycles of review and revisions implemented over time
Emergent aspects of validity
Improved understanding of the meaning of language assessment in context and of the effects and consequences on systems and people

of innovation suggests mechanisms which can be put in place to anticipate and achieve desirable outcomes through change processes. Fullan (1993:19), for example, suggests that the solution to achieving productive educational change 'lies in developing better ways of thinking about, and dealing with, inherently unpredictable processes'. His work also points up the social dimension of education and the relevance of theories of social systems and practices to assessment which have also been a focus of attention in language testing circles in recent years (e.g. McNamara & Roever 2006).

b. Socio-cognitive theories which place importance on both social and cognitive considerations are particularly relevant to the conceptualisation of language constructs (e.g. Weir 2005).

   The research methodologies needed to investigate the impact of examinations in their socio-cultural contexts indicate that insights from socio-cognitive theory might also be helpful in understanding how language learning and preparation for examinations takes place in formalised learning contexts. The literature on social psychology may also be relevant as social psychologists seek to explain human behaviour in terms of the interaction between mental state and social context; this is an important aspect of impact at the micro level.

c. Constructivism is important for the re-conceptualisation of impact for two reasons: first because contemporary approaches to teaching and learning in formal contexts now appeal to constructivist theories; second because it underpins the research paradigm which is most appropriate to finding out what goes on in contexts of test use, as seen in the case studies.

d. Contemporary theories of knowledge and of language learning need to play a more prominent role in the study of impact. For example, from the learner's perspective, affective factors are vital for motivation, and feedback from tests that highlights strengths positively tends to lead to better learning (assessment for learning).

These considerations are relevant in designing language assessment systems with learning-oriented objectives, and whether these objectives have been met is a concern in impact research.

Dimension 2: introduce the concept of 'impact by design' into the planning and operationalisation of language assessments by examination providers.

The concept of 'impact by design' is a key feature of the expanded impact model. This means designing tests which have the potential for positive impacts, including well-defined focal constructs supported by contemporary theories of communicative language ability, language acquisition and assessment (cf. the socio-cognitive model). It takes an *ex ante* approach to anticipating the possible consequences of a given policy 'before the event'.

'Impact by design' builds on Messick's idea (1996) of achieving '*validity by design as a basis for washback*'. The importance of the rational model of test development and validation with iterative cycles is a necessary condition for creating construct-valid tests and for the development of successful systems to support them.

At the heart of this is the adequate specification of the focal construct which is crucial for ensuring that the test is appropriate for its purpose and contexts of use (and to counter the twin threats to validity – construct under representation and construct-irrelevant variance – noted by Messick (1996: 252)).

This is a necessary condition for achieving the anticipated outcomes, but it is not sufficient and only provides the 'latent potential' for validity in use. For Cambridge ESOL impact by design highlights the importance of designing and implementing assessment systems, which extend the design features beyond the technical validities related to the construct, and incorporate considerations explicitly related to the social and educational contexts of test use.

As time passes following the introduction of an examination, new contexts of use arise and new users acquire a stake in the examination. As this extension of 'ownership' happens, there is a risk of 'drift' away from the original intentions of the test developers; for example, the intended relationship between use of test results and the test construct may begin to change over time due to influences in the wider educational context. The potential for negative impact is likely to increase when the original construct is no longer suitable for the decisions which the new users are making. In other words, the examination is no longer 'fit for purpose' and so corrective action of some kind needs to be taken.

Similarly, consequences – intended and unintended – often emerge after the test has been 'installed' into real-life contexts of use which are not uniform and are constantly changing as a result of localised socio-political and other factors. The overall validity of an assessment system, therefore, is *an emergent property* resulting from a test interacting with contexts over time.

'Impact by design' is therefore not strictly about prediction; a more appropriate term might be 'anticipation'. In working with stakeholders, possible impacts on both micro and macro levels can be anticipated as part of the design and development process. Where negative consequences are anticipated, potential remedial actions or mitigations can be planned in advance. So, for example, if 'construct drift' is a risk, it can be anticipated and appropriate tolerances set before test revisions are required. This approach is congruent with the concept of *social impact assessment*, a form of policy-oriented social research.

Dimension 3: re-organise validation procedures to incorporate impact research into operational activities to provide the basis for knowing about and understanding how well an assessment system works in practice with regard to its impact.

It is essential to know what happens when a test is introduced into its intended contexts of use; this should constitute a long-term validation plan, as required by the impact by design concept.

Finding out and understanding needs to be a routine preoccupation within the operational procedures and should be problematised within a research agenda which allows for impact-related research studies to be conducted where appropriate.

The emergentist approach noted above encourages impact researchers to develop an 'impact toolkit' of methods and approaches to 'finding out' (e.g. to carry out analyses of large-scale aggregated data, as well as micro-analyses of views, attitudes and behaviours in local settings). The quantitative analysis of macro-level group data can capture overall patterns and trends, while the qualitative analysis of multiple single cases enables the impact researchers to monitor variability in local settings and to work with the 'ecological' features of context.

While not rejecting experimental methods, an expanded model of impact looks to 'real world' research paradigms to provide tools which can shed light on what happens in testing contexts. Constructivist approaches to social research include mixed methods and quasi-experimental designs, as shown in the three cases reviewed in this thesis. Case studies are especially useful for investigating impact at the micro level and for understanding the complexities of interaction between macro-level policies and implementation in local settings. Without such methods it is difficult to find out about and understand how the interaction of differing beliefs and attitudes can lead to consensus or to divergence and diversity.

It is important for examination boards to modify their validation procedures in order to collect, store and access the necessary data and greater attention should be given to the planning and resourcing for this area of validation.

Dimension 4: develop an appropriate theory of action which enables examination providers to work with stakeholders to achieve the intended objectives, to avoid negative consequences and to take remedial action when necessary.

The ability to change systems to improve educational outcomes or mitigate negative consequences associated with the examinations is ultimately the most important dimension of the model. Anticipating impacts and finding out what happens in practice are not enough if improvements do not occur as a result; a theory of action is therefore required to guide practice.

Examples of theory of action are found in the literature on educational reform and school improvements, especially in the USA. Such examples provide support for the ways in which the four dimensions of the expanded model fit together in practice (e.g. Resnick and Glennan 2002). A theory of action provides planners and practitioners with the capacity to act in social contexts, to determine what needs to be done and when/how to do it. Being prepared to change and to manage change is critical to a theory of action. The challenge for the examination provider is to 'harness the forces of change' in order to get the relevant stakeholders working together to achieve better assessment outcomes.

Some of the dilemmas which arise in assessment contexts can only be dealt with if a wide range of stakeholders agrees to manage them in ways which they find acceptable. As Fullan (1999:xx) puts it: 'Top-down mandates and bottom-up energies need each other.'

Conclusion

The outcome of the thesis is an expanded model which was designed to help Cambridge ESOL and other examination providers to address the challenge of finding out and understanding how their examinations impact on society. Concrete and relevant applications for investigating the impact of language assessment at micro and macro levels within the routine work of the examinations board were also suggested.

References

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

Bachman, L (1990) *Fundamental considerations in language testing*, Oxford: Oxford University Press.

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

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

Cheng, L (1997) *The Washback Effect of Public Examination Change on Classroom Teaching: An impact study of the 1996 Hong Kong Certificate of Education in English on the classroom teaching of English in Hong Kong secondary schools*, unpublished PhD thesis, University of Hong Kong.

Cheng, L (2005) *Changing Language Teaching through Language Testing: A washback study*, Cambridge: Cambridge ESOL and Cambridge University Press.

Cheng, L and Watanabe, Y with Curtis, A (Eds) (2004) *Washback in language testing: Research contexts and methods*, Mahwah, NJ: Lawrence Erlbaum Associates.

Fullan, M (1991) *The New Meaning of Educational Change* (2nd ed.), London: Cassell.

Fullan, M (1993) *Change Forces: Probing the Depths of Educational Reform*, London: the Falmer Press.

Fullan, M (1999) *Change Forces: The Sequel*, London: the Falmer Press.

Green, A (2003) *Test Impact and EAP: a comparative study in backwash between IELTS preparation and university pre-sessional courses*, unpublished PhD thesis, the University of Surrey at Roehampton.

Green, A (2007) *IELTS Washback in Context; Preparation for academic writing in higher education*, Cambridge: Cambridge ESOL and Cambridge University Press.

Hawkey, R (2006) *The theory and practice of impact studies: Messages from studies of the IELTS test and Progetto Lingue 2000:* Cambridge ESOL/Cambridge University Press.

Henrichsen, L E (1989) *Diffusion of innovations in English language teaching: The ELEC effort in Japan, 1956–1968*, New York: Greenwood Press.

Hughes, A (1989) *Testing for language teachers*, Cambridge: Cambridge University Press.

McNamara, T and Roever, C (2006) *Language Testing: the Social Dimension*, Oxford: Blackwell.

Messick, S (1989) Validity, in Linn, R L (Ed) *Educational measurement* (3rd ed), New York: Macmillan, 13–103.

Messick, S (1996) Validity and washback in language testing, *Language Testing* 13 (3), 241–256.

Milanovic, M and Saville, N (1996) *Considering the Impact of Cambridge EFL Examinations*, Manuscript Internal Report, Cambridge: Cambridge ESOL.

Resnick, L B and Glennan, T K (2002) Leadership for learning: A theory of action for urban school districts, in Hightower, A M, Knapp, M S, Marsh, J A and McLaughlin, M W (Eds), *School Districts and Instructional Renewal*, New York: Teachers College Press, 160–172.

Saville, N (2003) The process of test development and revision within Cambridge EFL, in Weir, C and Milanovic, M (2003) (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Cambridge: Cambridge ESOL/Cambridge University Press.

Sayer, A (1984) Method in Social Science: *A Realist Approach*, Routledge: London.

Sayer, A (2000) *Realism and Social Science*, Sage: London.

Thelen, E and Smith, L B (1994) *A Dynamic Systems Approach to the Development of Cognition and Action*, Cambridge, MA: The MIT Press.

Van Geert, P (2007) Dynamic systems in second language learning: Some general methodological reflections, *Bilingualism: Language and Cognition* 10, 47–49.

Wall, D (1999) *The impact of high-stakes examinations on classroom teaching: a case study using insights from testing and innovation theory*, unpublished PhD thesis, Lancaster University.

Wall, D (2005) *The Impact of High-Stakes Testing on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory*, Cambridge: Cambridge ESOL and Cambridge University Press.

Watanabe, Y (1997) *The Washback Effects of the Japanese University Entrance Examinations of English – Classroom-based Research*, unpublished PhD thesis, University of Lancaster.

Watanabe, Y (2004) Teacher factors mediating washback, in Cheng, L and Watanabe, Y (Eds) with Curtis A, *Washback in language testing: Research contexts and methods*, Mahwah, N. J.: Lawrence Erlbaum Associates, 19–36.

Weir, C J (2005) *Language Testing and Validation: An Evidence-based Approach*, Basingstoke: Palgrave Macmillan.

Construct validation of the Reading module of an EAP proficiency test battery

HANAN KHALIFA RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

This summary is based on a doctoral thesis submitted to the University of Reading (UK) in 1997. The PhD was supervised by Professor Cyril Weir.

Research purpose

The research sought to establish the construct validity of the Reading module of an English for Academic Purposes (EAP) Graduate Proficiency Test (GPT) Battery developed by the ESP Center of Alexandria University in Egypt. It investigated the componential nature of the reading construct and the effect of background knowledge on test performance. Only full consideration of these two issues would substantiate validation of the Reading module.

Research questions

The Reading module of the Egyptian Graduate Proficiency Test Battery (GPT) was intended to measure global and local comprehension. In the study, 'global comprehension' refers to understanding propositions at the macro-structure level of the text and 'local comprehension' refers to understanding propositions at the micro-structure level. The former is concerned with the relationships between ideas represented in complexes of propositions or paragraphs which tend to be logical or rhetorical (see Vipond 1980), whereas the latter is concerned with the relationships between individual sentences or concepts which tend to be mechanical or syntactical. Reading at the global level involves skimming to establish the gist of the text, search reading to locate information on a pre-determined topic, and careful reading to understand explicitly and implicitly stated main ideas. Reading at the local comprehension level is operationalised by scanning to locate specific information, and reading carefully to infer the meaning of lexical items and identify pronominal referents. Global and local comprehension levels are characterised by two different rates of reading. Operations like skimming, search reading, and scanning require a faster reading rate than those involving careful reading at microlinguistic level. Weir & Urquhart (1998) refer to the former as *expeditious reading operations* (whereby the reader processes text quickly, selectively and efficiently) while they refer to the latter as *slow careful reading operations*.

On reviewing empirical evidence provided by product- and process-oriented studies, it became apparent that there is a case for and against the multi-divisible nature of reading. Product-oriented studies like that of Berkoff (1979), Carver (1992), Davis (1968), Guthrie & Kirsch (1987) and process-oriented studies (e.g. Anderson, Bachman, Perkins & Cohen 1991, Cohen 1984, Hosenfeld 1977, Nevo 1989) have provided empirical evidence for the separability of skills. On the other hand, product-oriented studies (e.g. Lunzer, Waite & Dolan 1979, Rosenshine 1980, Rost 1993, Thorndike 1973) and process-oriented studies like that of Alderson (1990a & b) have provided evidence that reading is a single holistic process. What is most significant in all of these studies is the occurrence of vocabulary as a second factor (also referred to as word meaning, verbal reasoning, word knowledge, semantic difficulty).

The contradiction in findings seemed to be due to sample selection and methodology used. First, process-oriented studies researched at that time highlighted the absence of a working definition of the operations used in tests, hence, disagreement among experts on what skill each item tested. Second, most of the product-oriented studies did not take into account the ability to process text quickly, i.e. the tests used do not exhibit a wide coverage of putative

EAP reading operations. Third, most of the studies favouring a unitary concept have been carried out on young learners and in L1 contexts. The case might, therefore, be different if the sample used were adult non-native speakers who are spread out across a range of language proficiencies. It may well be that for such a sample a distinction between lower-order skills and higher-order skills is valid (see Clarke 1980, Eskey & Grabe 1988).

The fact that the Reading module was intended to measure a variety of reading operations, and that some studies provided evidence for the emergence of certain operations as factors separate from a general reading competence one, provided the rationale for the formulation of the first research question.

> Research Question 1:
> Within which of the three Reading tests of the GPT Reading module are the components of these tests testing different reading operations as claimed by the module designers?

The starting point for the second research question was Weir & Porter's (1994) suggestion, based on reviewing some empirical data, that tests which include items testing local lower-order skills might discriminate against the micro-linguistically disadvantaged but otherwise competent reader. Similarly, Alderson & Lukmani's (1989) study has shown that weaker students tended to cope quite well with the text and questions at the global level but this was not matched by their performance on questions focusing on microlinguistic items at the local level. Thus, the researcher set out to investigate whether candidates were disadvantaged by the inclusion of any of the subtests, hence, the formulation of the second research question where group and individual performances are considered.

> Research Question 2:
> (A) Do groups at different levels of proficiency perform the same across the four components of each test in the GPT Reading module?
> (B) Do individuals perform the same across the four components of each test in the GPT Reading module?

The discussion of the nature of the reading construct posed another question: if sub-skills exist, do they interact with other factors such as text organisation or readers' familiarity with test content? It seemed quite obvious that drawing inferences can be easy when the reader has adequate background knowledge about the topic. When discussing reading comprehension, we cannot discuss just the interaction between the reader and the reading operations, but also the interaction between the reader and the text, in other words, the role of readers' background knowledge in text comprehension. Several studies (e.g. Alderson & Urquhart 1983, 1985 & 1988, Ausubel 1960, Clapham 1994 & 1996, Erickson & Molloy 1983, Ja'far 1992, Jensen & Hansen 1995, Kattan 1990, Koh 1985, Moy 1975, Peretz & Shoham 1990, Shoham, Peretz & Vorhaus 1987, Tan 1990) have investigated the effect of content familiarity on candidates' performance in EAP reading tests. Data emerging from these studies gives some tentative indication that there is a relation between candidates' background knowledge in their academic discipline and their performance on EAP reading comprehension tests.

When developing the specifications for the GPT, designers debated whether there should be separate academic modules for the disciplines involved. Ultimately, they decided that the Reading module would have texts covering three broad academic discipline areas: (1) Arts, Social Sciences, Administrative and Business Studies (ASAB); (2) Sciences (SS); and (3) Dentistry, Medicine, and Health Sciences (DMHS). This decision was based on three views. First, if one were to design discipline-specific modules for all disciplines it would clearly be a very large undertaking. Second, variation within a discipline area inevitably meant that one module was by no means specific for all the candidates doing that module. Third, there is as yet no body of evidence to support EAP testing claims that candidates are disadvantaged if they take a test which is not in the area of their discipline. The grouping of disciplines into three broad areas and classification of candidates accordingly were based on the lists supplied by the Student Affairs Divisions in Alexandria (Egypt) and Reading (UK) universities.

The third research question explored the value of including subject-specific reading tests in EAP testing. What is meant by subject specific is 'specific to the broad discipline areas', for example, specific to the area of Science disciplines.

> Research Question 3:
> Will postgraduate candidates in three broad discipline areas perform better on a Reading Comprehension test whose content is on a topic that is related to their own broad discipline area than on a Reading Comprehension test whose content is on a topic that is related to another broad discipline area, given that the texts are of approximately comparable difficulty?

Studies in ESP testing examined at the time also appeared to suggest that other factors are at play and that these factors seemed to be influencing the results or leading to conflicting results. We could divide these factors into two types: test-related factors, such as sample size, sample linguistic homogeneity, and sample academic level; and text-related factors, such as text specificity, text difficulty, and topic familiarity. Thus, the fourth research question attempted to find out which of these factors contributes most to candidates' performance on EAP Reading Comprehension tests.

> Research Question 4:
> Which contributes more to candidates' EAP reading proficiency scores: topic familiarity, topic/text ease, or L2 proficiency level?

Research methods

Quantitative and qualitative research methods were used to investigate the above research questions. This included: mindmapping, introspection procedures, feedback questionnaires and statistical analysis.

Instruments

To ensure that reading construct as defined by the test designers was adequately captured by the test items, the items were matched against mindmaps of the text produced by subject and language experts. This procedure was used to justify the existence of the test items, and to re-categorise the items under four subtests. A panel of language and subject experts was asked to provide mindmaps of the texts and identify key lexical words. They went through the operations the items were supposed to test. A synthesis of information was then collected from the mindmaps. Items which did not feature in this consensus or on which expert judges widely disagreed were marked for possible exclusion from the tests.

The mindmapping procedure was followed by an introspection activity. The first part of this activity was used to establish whether each item measured what it was designed to measure. Another group of language and subject experts and a group of proficient subject students were asked to introspect on what skill(s) they use in answering the items. The second part of the activity consisted of retrospection interviews with subject students. Interviews were conducted to clarify those cases where candidates had arrived at the same response via a process different from the expected one, and to ask why candidates had left an item unanswered or had used more than one skill. The introspection procedure was a way of gaining insights into how readers arrive at their answers and of determining if test items were testing what they claimed to test.

The module was then administered and data was subjected to classical and rasch analyses. Decisions on which items to exclude or retain depended on the pulling of evidence from three different data sources: meaning and lexical consensus, introspection proforma, and item analysis.

In order to investigate research questions 2(A) and 4, it was necessary to have a common measure of proficiency so that candidates could be placed into language levels. Thus, a vocabulary and grammar test which was part of the Test of English for Educational Purposes (TEEP) (see Weir 1988) was used. Candidates were divided into three levels in accordance with Egyptian universities' proficiency level requirements for admission to postgraduate courses.

In order to investigate research question 4, two sets of questionnaires were used to find out about text specificity, topic/text ease, and topic familiarity. The subject lecturers' questionnaire was used to find out how they assessed the specificity, familiarity, and difficulty of the Reading module texts on a 4-point scale (high, medium, low, not at all) according to their knowledge of their students' level of proficiency and of the discipline knowledge they thought their students might use in answering the items. The term 'specific' here was used to indicate how specific the topic was, how specific the vocabulary used in the text, and how specific the non-linear information given in the text were to their postgraduate students. Familiarity was defined in terms of the topic and the rhetorical organisation of the texts and tasks required to answer the test items. Difficulty was seen in terms of language in a text and item difficulty.

The test takers' feedback questionnaire was used to find out about perceived topic familiarity, and perceived topic ease/test bias. A 3-point scale was used for those items. It should be pointed out that the questionnaires were administered to test takers immediately after they had finished the tests. Since candidates did not take any two tests immediately after each other, there is no reason to believe that in answering the questionnaires candidates were comparing texts.

Participants

Candidates who participated in this research comprised two sub-samples: linguistically heterogeneous and linguistically homogeneous EAP learners. The homogeneous sample consisted of 973 non-native speakers of English registering for postgraduate courses at Alexandria University in Egypt. Candidates here share the same L1 background (i.e. Arabic). They were classified into the three broad discipline areas described above. The heterogeneous sample consisted of 355 non-native speakers of English. These were registering for postgraduate courses at Reading University in England. Candidates in this sample had different L1s (e.g. Chinese, French, Japanese, Danish, Italian, Turkish). Candidates were classified into two broad discipline areas: Arts and Sciences. There is no Medical group in the UK sample since Reading University does not provide courses for candidates in this group.

Forty-five subject lecturers (of near native proficiency in English) who were teaching postgraduates in Alexandria University in Egypt participated in the study. Lecturers in Arts disciplines were teaching at the faculties of Arts, Fine Arts, Commerce and Tourism. Science disciplines lecturers were teaching at Agriculture, Engineering and Science faculties. Lecturers from the Medical disciplines were teaching at the faculties of Dentistry, Medicine, Nursing and Pharmacy. No data was collected from subject lecturers in the UK due to practical constraints.

Results and discussion

Research question 1

In order to investigate the first research question, qualitative data from introspection proforma and retrospection interviews as well as quantitative data from subtests' inter-correlations and factor analysis were collected from Egyptian and UK pre-sessional samples taking a single test: the Arts Test, the Science Test, or the Medicine Test.

All three tests exhibited low inter-correlations between subtests measuring global and local comprehension, and between subtests requiring expeditious and careful reading. Factor analysis gave an indication that the tests were not operating uni-dimensionally. It showed the consistent presence of at least a second factor. It also appeared to suggest that candidates behave differently on the operations being tested: a clear factor structure showing a distinction between expeditious and careful reading occurred across a range of samples of EAP candidates taking different tests. This is in line with Guthrie & Kirsch's (1987) and Carver's (1992) findings that made a case for differentiating between reading to comprehend explicitly stated ideas and reading to locate specific information.

Similarly, introspection proforma and retrospective interviews indicated that the operations the subject students reported using to answer the test items differed according to the subtest they were answering. For example, in the scanning subtest, students reported rapid inspection of the text; going backward and forward in the text looking for specific words, dates, etc. In contrast, in the reading carefully subtest students reported slow inspection of the text; observance of the linearity and sequencing of the text. They read and reread in order to establish more clearly and accurately the comprehension of main ideas.

On the whole, the answer to this research question is 'Yes'. Findings from qualitative and quantitative research methods appear to support the test designers' claim that the tests are measuring separable subskills, and lend support to the argument for the existence of separate reading operations. They, therefore, contradict the oft-expressed view that reading is a unitary construct.

Research question 2

For research question 2, group and individual performances in the linguistically homogeneous and heterogeneous samples in single and paired data sets were looked at. The Grammar Test was used as a measure of candidates' general language ability and to classify them into high, middle, and low proficiency level groups. Cross-tabulations were used. The intention was to compare the performances of individuals who passed and those who failed in each of the GPT Reading module tests. Research findings provided evidence for significant differential performance on the components of the tests.

In most cases candidates perform better on global items than on local items. This seems to be in line with the findings of Alderson & Lukmani's (1989) study. Similarly, most of the evidence shows that candidates of different ability levels seem to perform better on items requiring slow careful reading than those requiring expeditious reading. This is in line with Beard (1972) and Weir (1983) whose studies into students' abilities indicate that 'for many readers reading quickly and efficiently posed greater problems than reading carefully and efficiently' (Weir 1998). This draws attention to Weir & Urquhart's (1998) call for 'paying attention to expeditious reading strategies in both teaching and testing'. It should be noted that candidates of different proficiency levels performed the worst on scanning, with the low-level groups being the most severely disadvantaged by the inclusion of scanning items in a Reading Comprehension test.

The results of cross-tabulations for individual performances affirmed those reported for the group data. The most interesting finding, however, came from the paired data sets. These showed that, across two tests, not a single individual performed consistently better on local than on global comprehension components, or on expeditious than on slow careful reading components. In contrast, the results showed a number of individuals who consistently passed on global and failed on the local comprehension parts of both tests, and others who consistently passed on slow careful but failed on the expeditious reading parts.

Overall, the findings indicate that candidates perform differentially on the subtests. Some appear to be disadvantaged by the expeditious reading subtests compared to the careful reading ones, while others appear to be disadvantaged by the local comprehension subtests compared to the global ones. In certain individuals, however, the case may be more marked in terms of global and local or expeditious and slow. Individuals vary in their profile of proficiency – where local comprehension might be weaker than global comprehension and expeditious reading weaker than slow reading, for instance. Furthermore, these differences may vary considerably with level of candidates and according to text. It is clear from this data that a serious case can be made for the profiling of abilities in each of the skill operations; otherwise false conclusions may be drawn about candidates' reading ability.

Research question 3

There seemed to be no straightforward answer to this research question. The findings showed that the evidence is mixed. For the entire test population, no significant difference was observed between the performance of the different discipline groups. Candidates did not seem to either suffer or profit from taking Reading tests in different discipline areas. This finding is compatible with those of Carrell (1983) and Clapham (1993, 1994, 1996).

When looking at group performances in the paired data sets, significant differences were found. Both discipline groups (Arts and Sciences) of the linguistically heterogeneous sample appeared to suffer when taking the Science Test and profit when taking the Arts Test. In contrast, each of the three groups of the linguistically homogeneous sample (Arts, Sciences, Medicine) appeared to be at an advantage when taking the Science Test and at a disadvantage when taking the Arts Test. This picture was confirmed when considering individual performances in the paired data sets.

In considering the findings of this research question, it should be noted that the value of using a homogeneous sample is that candidates share the same L1, similar instructional background, or previous learning experiences, that is, variables that were not controlled for in the heterogeneous sample and might have neutralised the subject effect for this sample. It should also be noted that the texts in the GPT Reading module were selected from academic journals in the appropriate broad discipline areas. They were expected to be appropriate and specific to the relevant Reading module and, therefore, by implication to be unsuitable for or unfamiliar to candidates in other disciplines. However, the evidence provided by this research showed that, in some cases, this is not necessarily the case. One possible explanation could be that studying in one particular discipline area does not mean that candidates are ignorant about other disciplines or unfamiliar with other rhetorical structures. They may well read books and articles in disciplines outside their own academic field.

The findings of the third research question seem to indicate that if there is to be one test catering for candidates from different disciplines for reasons of practicality, then there is evidence coming out of the heterogeneous sample to suggest the selection of a humanities-based test. However, there is evidence provided by the homogeneous sample to suggest that candidates would suffer if they take a humanities-based test and profit if they take a science-based one. This implies that the argument for ESP testing remains unproven at the time the study was conducted.

Research question 4

In order to investigate this research question, subject lecturers' and test takers' views on topic familiarity and on topic/text ease were considered. Here, only the paired data set of the Egyptian sample was used. A multiple linear regression analysis was also conducted to find out what proportion of the total variability in candidates' scores can be explained by proficiency level, topic familiarity, and topic/text ease, as well as which combination of these variables most accurately predicts candidates' performance or which do not add much to the prediction. The Grammar Test was used as a way of determining proficiency levels.

There is some evidence to suggest that subject lecturers' views on topic familiarity, on the whole, seem to be a good indicator of candidates' test performance. For example, when subject lecturers said their students were more familiar with the Science Test topic than with the Arts Test topic, this was reflected in the students' better performance on the Science Test. Thus, in selecting texts it seems worthwhile to collect such views in order to reduce test bias by eliminating unfamiliar test topics. There is also some evidence to suggest that test takers' subjective evaluation of the relative difficulty of a Reading text (as measured by their views on topic/text ease) is not always a good indicator of their actual performance on Reading Comprehension tests. This finding is consistent with Carrell's suggestion that 'non-native readers appear not to have good sense of how easy or difficult a text is for them to understand' (Carrell 1983:183).

The results of multiple regression analysis, which was run separately on the Arts and Science Tests, showed that topic familiarity and proficiency level contributed the most to candidates' reading proficiency scores. Proficiency level appeared to have a much stronger effect on candidates' scores on the Arts Test than did topic familiarity, whereas it was the other way round in the Science Test. This seems to imply the existence of text effect. One can only speculate on the reasons, though. It might be possible that candidates resorted to their knowledge of the language because they found the Arts text more difficult than the Science one. On the other hand, it might be possible that candidates found the Science text more specific than the Arts one so they turned to their knowledge of the topic and capitalised on their familiarity with the topic. The contribution of topic/text ease was the least marked. In fact this variable did not contribute much to the regression equation. These findings are compatible with Mohammed & Swales (1984) and Zuck & Zuck (1984) who found that topic familiarity is often a greater predictor of comprehension ability than are text-based linguistic factors such as syntactic ease.

Conclusion

Separability of reading operations

It seemed that, irrespective of texts or item difficulty, differential performances on the components of the Reading tests occurred. This has implications for teaching and testing EAP reading at least within the Egyptian context.

Firstly, the findings of this study seemed to suggest that training in expeditious reading strategies may still be inadequate in the EAP classroom. It seems that the tendency in these classes is to focus on careful reading, while expeditious reading in the sense of efficient and quick reading seems to be neglected to a large extent. If the aim is to have efficient readers, then the teaching tasks should also include the practice of expeditious reading operations. If students are trained to use various expeditious operations to deal with different reading tasks, they will be able to cope with similar real world reading tasks better. As Weir & Urquhart (1998) suggest, different passages could be used for teaching expeditious and careful reading to make students aware of the flexibility of using different approaches to different texts and different tasks.

Secondly, although expeditious reading operations have been incorporated into reading materials, they have been, to a large extent, overlooked by test designers whose focus at the time of conducting the study was mainly on careful reading. The study, with its literature review and data analyses drawn from various sources, reflected the need to include a subtest forming expeditious reading operations in EAP tests. Similarly, in view of construct validity test designers can hardly ignore such a need.

Thirdly, the profiling of abilities seemed fairer than reporting results as a composite score. In other words, in a case like the GPT reading tests, the aim should be to produce a profile of: the ability to read expeditiously at the global level, the ability to read expeditiously at the local level, the ability to read carefully at the global level, and the ability to read carefully at the local level.

Number of texts

In terms of task effect, the evidence provided by the present study forces us to accept that, despite rigorous test development procedures, item/component difficulty may vary from one test to another; and that some tests will simply be easier to access than others. This might be due to factors like rhetorical organisation, macro-structure, and so on. The only real solution, therefore, is to develop clearer procedures for identifying these factors, or to use a range of texts for testing each component.

In terms of text effect, the evidence supplied is mixed and it appears that the nature of the research sample is playing an important role. For example, the UK pre-sessional data suggests that if there has to be one text, a humanities-based one would be the least disadvantageous. However, data from the Egyptian homogeneous sample suggests that candidates would be better off with a science text. One can only speculate that this mixed evidence might be due to differences in instructional background. There seems to be a need, therefore, to consider carefully candidates' previous education when deciding on the number and nature of texts to be used in an EAP Reading test.

Given the above mixed evidence, it would probably be safer to use a variety of texts from the broad discipline areas. If either one text or more than one is opted for, there should be systematic ways followed in text selection. The next section describes the basis on which texts should be selected.

Selection of texts

In selecting texts, the importance of face validity cannot be ignored if just one academic Reading test is opted for. We cannot ignore what subject lecturers and test takers said regarding text specificity and topic familiarity. In addition, it might be hard to get an approval from university authorities to use a test which seemingly does not look subject specific. The danger also exists that under examination conditions some students might be upset by the apparently unfamiliar material. In turn, they might not do as well as they should have.

On the other hand, if there is a need to create parallel EAP Reading tests, it seems quite impossible to find texts which are similar in terms of specificity, difficulty, and familiarity unless either a general academic text or a very highly specific one is opted for. If the latter is chosen, then the number of candidates who would be sitting for such a test is inevitably limited. In addition, tests which are too specialised may assess subject matter knowledge in a particular field more than the reading ability of the candidates, and thus individuals who happen to have less subject matter knowledge might be discriminated against. Thus one is forced to choose texts which are equally comprehensible for, and generally accessible to candidates in all fields within the broad discipline areas. They should come from an academic source and have an academic nature. The rhetorical structure could be argumentative or Introduction-Methods-Results-Discussion (IMRD), the former being more suitable to humanities-oriented candidates and the latter to scientifically-oriented candidates.

In other words, in developing Reading tests which cater for a large number of candidates, there is a need to ensure that the chosen topic is fairly familiar to all candidates so as to avoid bias caused by topic familiarity. Several texts of different topics might be used to counter-balance the topic-familiarity effect. The level of difficulty of the test should also be taken into account. The texts also would have to be submitted to subject specialists and students to check that no discipline is advantaged over another. These factors appear to be crucial to test designers to get stable, reliable, and meaningful results. Thus what seems to be needed is the development of a mechanism to screen texts for difficulty, familiarity, and specificity.

Triangulation of data sources

In empirically validating the GPT Reading module, information was collected from a variety of sources: experts' mindmapping consensus, experts' and subject students' introspection proforma, subject students' retrospection interviews, and item statistical analyses. Following such a procedure provided a sound basis for the final version of the tests in the Reading module. The mindmapping consensus eliminated idiosyncrasies that existed in content selection. The introspection proforma appeared to enhance the probability that the required operations were being tested. The retrospection interviews illuminated, to some extent, how the behaviour test items produced may equate with the behaviour identified in the theory-based model.

References

Alderson, J C (1990a) Testing Reading Comprehension Skills (Part One), *Reading in a Foreign Language* 6 (2), 425–438.

Alderson, J C (1990b) Testing Reading Comprehension Skills: Getting Students to Talk about Taking a Reading Test (Part Two), *Reading in a Foreign Language* 7 (1), 465–503.

Alderson, J C and Lukmani, Y (1989) Cognition and Reading: Cognitive Levels as Embodied in Test Questions, *Reading in a Foreign Language* 5 (2), 253–270.

Alderson, J C and Urquhart, A H (1983) The Effect of Students' Background Discipline on Comprehension: a Pilot Study, in Hughes, A and Porter, D (Eds) *Current Developments in Language Testing*, London: Academic Press, 121–138.

Alderson, J C and Urquhart, A H (1985) The Effect of Students' Academic Discipline on Their Performance on ESP Reading Tests, *Language Testing* 2 (2), 192–204.

Alderson, J C and Urquhart, A H (1988) This Test is Unfair: I'm not an Economist, in Carrell, P L, Devine, J and Eskey, D E (Eds) *Interactive Approaches to Second Language Reading*, Cambridge: Cambridge University Press, 168–183.

Anderson, N J, Bachman, L, Perkins, K and Cohen, A (1991) An Exploratory Study into the Construct Validity of a Reading Comprehension Test: Triangulation of Data Sources, *Language Testing* 8 (1), 41–66.

Ausubel, D P (1960) The Use of Advance Organisers in the Learning and Retention of Meaning Material, *Journal of Educational Psychology* 51, 267–272.

Beard, R (1972) *Teaching and Learning in Higher Education*, Harmondsworth: Penguin Books Ltd.

Berkoff, N A (1979) Reading Skills in Extended Discourse in English as a Foreign Language, *Journal of Research in Reading* 2 (2), 95–107.

Carrell, P L (1983) Some Issues in Studying the Role of Schemata or Background Knowledge in Second Language Comprehension, *Reading in a Foreign Language* 1 (2), 81–92.

Carver, R P (1992) Reading Rate: Theory, Research and Practical Implications, *Journal of Reading* 36 (2), 84–95.

Clapham, C (1993) Is ESP Justified? in Douglas, D and Chapelle, C (Eds) *A New Decade of Language Testing Research*, TESOL, 257–271.

Clapham, C (1994) *The Effect of Background Knowledge on EAP Reading Test Performance*, unpublished PhD thesis, University of Lancaster.

Clapham, C (1996) *The Development of IELTS: A Study into the Effect of Background Knowledge on Reading Comprehension*, Cambridge: University of Cambridge Local Examinations Syndicate.

Clarke, M A (1980) The Short-circuit Hypothesis of ESL Reading – or When Language Competence Interferes with Reading Performance, *The Modern Language Journal* 64 (2), 104–109.

Cohen, A D (1984) On Taking Tests: What the Students Report, *Language Testing* 1 (1), 70–81.

Davis, F B (1968) Research in Comprehension in Reading, *Reading Research Quarterly* 3 (4), 499–545.

Erickson, M and Molloy, J (1983) ESP Test Development for Engineering Students, in Oller, J W (Ed.) *Issues in Language Testing Research*, Rowley, Mass.: Newbury House, 280–300.

Eskey, D E and Grabe, W (1988) Interactive Models for Second Language Reading: Perspectives on Instruction, in Carrell, P L, Devine, J and Eskey, D E (Eds) *Interactive Approaches to Second Language Reading*, Cambridge: Cambridge University Press, 223–236.

Guthrie, J T and Kirsch, I S (1987) Distinctions Between Reading Comprehension and Locating Information in Text, *Journal of Educational Psychology* 79, 220–228.

Hosenfeld, C (1977) A Preliminary Investigation of the Reading Strategies of Successful and Nonsuccessful Second Language Learners, *System* 5 (2), 110–123.

Ja'far, W M (1992) *The Interactive Effects of Background Knowledge on ESP Reading Comprehension Proficiency Tests*, unpublished PhD thesis, University of Reading.

Jensen, C and Hansen, C (1995) The Effect of Prior Knowledge on EAP Listening Test Performance, *Language Testing* 12 (1), 99–119.

Kattan, J (1990) *The Construction and Validation of an EAP Test for Second Year English and Nursing Majors at Bethlehem University*, unpublished PhD thesis, University of Lancaster.

Koh, M Y (1985) The Role of Prior Knowledge in Reading Comprehension, *Reading in a Foreign Language* 3 (1), 375–380.

Lunzer, E, Waite, M and Dolan, T (1979) Comprehension and Comprehension Tests, in Lunzer, E and Gardner, K (Eds) *The Effective Use of Reading*, London: Heinemann Educational, 37–71.

Mohammed, M A H and Swales, J M (1984) Factors Affecting the Successful Reading of Technical Instructions, *Reading in a Foreign Language* 2 (2), 206–217.

Moy, R (1975) *The Effect of Vocabulary Clues, Content Familiarity and English Proficiency on Cloze Scores*, unpublished PhD thesis, University of California.

Nevo, N (1989) Test-taking Strategies on a Multiple-choice Test of Reading Comprehension, *Language Testing* 6 (2), 199–215.

Peretz, A S and Shoham, M (1990) Testing Reading Comprehension in LSP: Does Topic Familiarity Affect Assessed Difficulty and Actual Performance?, *Reading in a Foreign Language* 7 (1), 447–455.

Rosenshine, B V (1980) Skill Hierarchies in Reading Comprehension, in Spiro, R J, Bruce, B C and Brewer, W F (Eds) *Theoretical Issues in Reading Comprehension*, Hillsdale, NJ: Erlbaum, 535–554.

Rost, D H (1993) Assessing the Different Components of Reading Comprehension: Fact or Fiction? *Language Testing* 10 (1), 79–92.

Shoham, M, Peretz, A S and Vorhaus, R (1987) Reading Comprehension Tests: General or Subject Specific?, *System* 15 (1), 81–88.

Tan, S H (1990) The Role of Prior Knowledge and Language Proficiency as Predictors of Reading Comprehension among Undergraduates, in de Jong, J H A L and Stevenson, D K (Eds), *Individualising the Assessment of Language Abilities*, Clevedon, PA: Multilingual Matters, 214–224.

Thorndike, R L (1973) Reading as Reasoning, *Reading Research Quarterly* 9, 135–147.

Vipond, D (1980) Micro- and Macro-processes in Text Comprehension, *Journal of Verbal Learning and Verbal Behaviour* 19, 276–296.

Weir, C J (1983) *Identifying the Language Problems of Overseas Students in Tertiary Education in the United Kingdom*, unpublished PhD thesis, University of London.

Weir, C J (1988) The Specification, Realisation and Validation of an English Language Proficiency Test, *ELT Documents: 127*, Modern English Publications: The British Council.

Weir, C J (1998) The Testing of Reading in a Second Language, *Language Testing & Assessment 7*, Kluwer: Dordrecht.

Weir, C J and Porter, D (1994) The Multi-Divisible or Unitary Nature of Reading: The Language Tester between Scylla and Charybdis, *Reading in a Foreign Language* 10 (2), 1–19.

Weir, C J and Urquhart, A H (1998) *Reading in a Second Language: Process and Product*, Longman.

Zuck, L V and Zuck, J G (1984) The Main Idea: Specialists and Non-specialist Judgements, in Pugh, A K and Ulijn, J M (Eds), *Reading for Professional Purposes: Studies and Practices in Native and Foreign Languages*, London: Heinemann Educational, 130–145.

Comparing proficiency levels in a multi-lingual assessment context

KAREN ASHTON RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

This short summary is based on a doctoral thesis submitted to the Faculty of Education, Cambridge University (UK) in 2008. The research was funded by Cambridge ESOL. The PhD was supervised by Dr Neil Jones and Dr Edith Esch.

The PhD research focused on Cambridge ESOL's Asset Languages assessments.

This mixed-methods PhD explores and compares the reading proficiency of secondary school learners of German, Japanese and Urdu in England with the aim of investigating and shedding light upon the feasibility of relating learners of different languages and contexts to the same framework. This research has important implications within education, particularly given the use of frameworks such as the National Curriculum for Modern Foreign Languages (DfES and QCA 1999) and the increasing use of the Common European Framework of Reference (CEFR hereafter) (Council of Europe 2001) both within England and Europe.

'Can Do' statements are commonly used, and are being promoted for wider adoption (see Council of Europe 2008), in educational assessment to describe the level of a learner's reading proficiency. However, there is no research as to how, or whether, such 'Can Do' frameworks can be applied to all languages, particularly non-Latin script or community languages. The majority of research in this area has focused on learners of English, although the few single language research studies undertaken indicate that reading in languages like Japanese and Urdu requires different processing strategies from reading in alphabetic languages such as German for learners with English as their first language. Existing research has also failed to relate findings to proficiency level, making it impossible to compare findings across studies.

This thesis employed a mixed-methods approach, using self-assessment 'Can Do' surveys and think-aloud protocols, to compare the reading proficiency of secondary school learners of German, Japanese and Urdu in England. Findings show that statistically the same three factors best represent learners' understanding of reading proficiency across all three languages. However, there are also strong differences. For example, the difficulty of script acquisition in Japanese impacts on learners' understanding of the construct, while learners of both Japanese and Urdu were unable to scan texts in the way learners of German were able to. Urdu learners under-rated their ability, not taking into account the wide range of natural contexts in which they use Urdu outside the classroom. The findings also illustrate how Urdu learners use their spoken knowledge of Urdu as a resource when reading. Finally, this research demonstrates that the construct of reading in the National Curriculum for Modern Foreign Languages is not endorsed by any of the learner groups, which is worrying for language education and assessment within England and raises the need for further research.

References

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

Council of Europe (2008) *Recommendations of the Committee of Ministers to member states on the use of the Common European Framework of Reference for Languages (CEFR) and the promotion of plurilingualism*, Strasbourg, Adopted by the Committee of Ministers on 2 July 2008.

DfES and QCA (1999) *The National Curriculum for England: Modern Foreign Languages*, London: DfEE/QCA.

Testing financial English: Specificity and appropriacy of purpose in ICFE

ANGELA WRIGHT BUSINESS MANAGEMENT GROUP, CAMBRIDGE ESOL

This short summary is based on a Master's thesis submitted to Anglia Ruskin University in 2007. The research was funded by Cambridge ESOL.

Developers of tests of languages for specific purposes are faced with the challenge of creating tests which allow for an appropriate interaction between subject knowledge and language ability in relation to the target language use domain. This dissertation was completed while the International Certificate in Financial English (ICFE) was under development and set out to establish the extent to which the Reading paper meets this challenge. The research aimed to establish the degree of specificity of the ICFE Reading paper, to try and identify the characteristics that make it specific, and to find out how appropriate it is as a testing instrument for people working in or intending to work in the financial domain. There were three stages in this research, each comparing ICFE to tests of General and Business English at the same level (CEFR levels B2/C1). In the first stage, a questionnaire was administered to both subject specialists and non-specialists. It was designed to measure the subject specificity and appropriacy of the texts used in ICFE. In the second stage, a questionnaire was given to testing specialists only. It was designed to measure the degree of specificity of various aspects of context validity in ICFE in comparison to Business and General tests. The third stage involved a corpus study which aimed to identify some of the characteristics of the core language of Financial English, by comparing Financial English texts to Business and General English texts. The results taken together suggest that ICFE might be placed at the more specific end of the 'specificity continuum' than the General and Business English tests, and that although there is considerable fuzziness between Financial and Business English, distinct linguistic differences were found between Financial and General English and the beginning of a core Financial lexis was identified. It was found that the degree of specificity of ICFE made it appropriate as a testing instrument in relation to the target domain. For more details on one of the aspects of this study see Wright (2008).

References

Wright, A (2008) A corpus-informed study of specificity in Financial English: the case of ICFE Reading, *Research Notes* 31, 16–21.

The expression of affect in spoken English

MARK ELLIOTT ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

This paper is based on a Master's thesis submitted to King's College London (UK) in 2008. The thesis was supervised by Susan Maingay and Dr Nick Andon.

When we speak, we do not merely transfer information from one individual to another; we also give expression to a whole range of emotions, attitudes and evaluations. This phenomenon, 'pervasive, because no text or utterance is ever absolutely free from it [and] elusive, because it may be difficult to say exactly what it is that gives the text or utterance that certain quality' (Dossena & Jucker 2007:7), is known as *affect*.

At present, affect tends to sit on the periphery of models of language and language proficiency, treated as an 'optional overlay of emotion' (Thompson & Hunston 2000:20) to the expression of 'core' informational meaning.

Affect can be broken down into two core areas: emotion and attitudes. Emotion covers feelings such as anger and happiness, while attitudes are an individual's opinions of the world, formed through predisposition, experience and ideology, and which colour his or her perceptions. Attitudes are realised in language by *evaluation* (Thompson & Hunston 2000), which are essentially good or bad value judgements. Evaluation 'does not occur in discrete items but can be identified across whole phrases, or units of meaning, and ... is cumulative' (Hunston 2007:39).

Affect can be expressed towards many different objects. These are most likely to be previous utterances, the proposition being made, agents implicated within the proposition, the listener or the speaker; there could, however, be still more.

Many different resources are employed in the expression of affect, and they interact in complex and sometimes unpredictable ways. To reflect this, this study is grounded in a *complex systems* view of language (Larsen-Freeman & Cameron 2008). The study considers how different elements of language interact within a specific context to create affective meaning.

Complex systems theory and language

'Tidy explanations survive as long as all that has to be explained is the meaning of sentences invented by armchair linguists' (Coates 1990:62).

Coates captures one of the tensions at the heart of applied linguistics. By focusing on small, manageable areas of the language and producing clear, tidy explanations, we can lose sight of the fact that real-life language simply does not behave in this fashion. In reality, the production of meaning is a highly complex process involving the interaction of a variety of components: lexis, grammar, phonology, discourse-level features, paralinguistic and non-verbal features and, crucially, context. Indeed, language exhibits many, if not all, of the properties of a *complex dynamic system* (Larsen-Freeman & Cameron 2008), and treating it as such provides a suitable framework for investigating the expression of affect.

Complex systems involve a large number of components interacting, often in a non-linear fashion (i.e. when a change in input results in a disproportionate change in output). Complex systems exhibit certain key features. Let us consider these, following Larsen-Freeman & Cameron (2008), with examples of how they relate to language:

1. **Heterogeneity of elements and agents:** the elements or agents in a complex system are often extremely diverse, and can be *processes* rather than entities, or even complex subsystems. Although the components may be diverse, they are interconnected – change in one component affects others. Language elements include phonetic and phonological features, lexis, grammar and discourse-level features; agents include users of the language (at an individual level) and society (at a higher level).
2. **Dynamics:** complex systems are in a permanent state of flux. Change takes place on *scales* (time) and *levels* (size): change may occur at the level of the whole system, a subsystem within it, or only a very small part of it. Different levels and scales influence each other upwards and downwards. Languages change on both micro levels (such as the introduction of a new word) and macro levels (such as changes in the formation of tenses), and both over short and long scales.
3. **Non-linearity:** due to the interconnected nature of the elements in a complex system, change can result which is out of proportion to the external stimulus. An example of this is the famous 'butterfly effect' (weather is an example of a complex system). Some language innovations spread rapidly through a language while others are ignored. Similarly, a slight change of intonation could render a completely different interpretation to an utterance.
4. **Openness:** complex systems are open. They can – and must – take on new elements and energy in order to remain in a state of *dynamic stability*, where the system is stable but not static or fixed. New words are constantly being created, either to label new developments in society and the world (the source of external energy), or from other languages through 'borrowed' words.
5. **Adaptation:** many complex systems are adaptive, meaning that change in one part of the system leads to change in the system as a whole, as it adapts to the new situation. Although languages are in constant flux, the basic requirement of intelligibility dictates that the language incorporates changes by adapting to new circumstances without losing its overall integrity.
6. **The importance of context:** context is crucial when considering complex systems – indeed, the context

within which a system operates cannot be considered separate from the system itself; it actually forms a part of the system. For example, no utterance in any language can be fully interpreted without consideration of the context it was uttered in, such as who uttered it, to whom and in what situation.

7. **Constructions**: *construction grammar* (Goldberg 2003) provides a model of grammar which is consistent with complex systems theory, and within which we shall frame this study. Constructions range from morphemes through words and chunks, up to abstract grammatical structures. Constructions carry inherent semantic or discoursal functions, rather than being 'empty' syntactic shells for meaning-carrying words. These semantic meanings can change over time – for example the *be going to* construction originally only denoted movement: *I'm going to the shops* (literally), but developed its present future meaning, as in: *I'm going to buy some bread there* (Perez 1990).

Discourse and complex systems

We try to understand language in use 'by looking at what the speaker says against the background of what he might have said but did not, as an actual in the *environment* of a potential' (Halliday 1978:52). This Systemic Functional viewpoint is echoed in a complex systems approach, where discourse is 'action in complex dynamic systems nested around the microgenetic moment of language using' (Larsen-Freeman & Cameron 2008:163). Individuals adapt their utterances to take into account all relevant contextual features.

In discourse, different scales and levels interact to create complex systems phenomena we have already encountered: self-organisation (the progression of the discourse), emergence (of meaning and new semiotic entities within the discourse) and reciprocal causality (between the interlocutors, and between the speakers and the discourse itself). The expression of affective meaning can be viewed as an *emergent phenomenon* from the interaction of the elements and agents of the complex system of discourse.

Affective resources

Speakers use a range of resources within the language to create affective meaning: lexis, grammar, phonology, discourse-level features and context. We will term these *affective resources*, and consider them in turn.

Lexis

Individual lexemes

Some words and phrases serve purely affective functions; *brilliant*, for example, has no ideational meaning beyond the evaluative. However, the affective meaning of an *utterance* is not determined by lexis alone. The utterance *That was brilliant* could convey its 'natural' semantic meaning, but in a different context and with sarcastic intonation, it could also convey precisely the opposite meaning. As Vološinov (1986:68) notes regarding the malleability of language: 'What is important for the speaker about a linguistic form is not that it is always a stable and self-equivalent signal, but that it is an always changeable and adaptable sign.' This is not to negate the importance of lexis, but merely to underline that it is one of several affective resources employed in an utterance; this holds true of all affective resources. In analysing a text, we need to consider the interaction of the affective resources.

There are other lexemes which encode ideational meanings whilst also expressing an affective *connotation*; these often exist in apposition to more affectively neutral alternatives. For example, the words *dog*, *doggie*, *cur* and *mutt* all have the same ideational referent, but encode rather different affective connotations.

Semantic prosody

A form of connotation can exist at another level through *semantic prosody* – how 'a given word or phrase may occur most frequently in the context of other words or phrases which are predominantly positive or negative in their evaluative orientation' (Channell 2000:38). In this way, connotations of collocants are 'inherited' by the word or phrase, often lending them an affective meaning which can develop across a text or texts. Corpus analysis of semantic prosodies has produced some interesting, not always intuitive, results – the phrase *par for the course*, for example, almost exclusively appears in cases of negative evaluation, so although it may not directly encode a negative connotation, it carries a negative semantic prosody (ibid.).

Grammar

Affective constructions

Wierzbicka (1987) argues that certain constructions encode specific affective meanings that cannot be accounted for by reference to conversational implicature alone. I will term such constructions, which encode an affective meaning either instead of or in addition to an ideational meaning, *affective constructions*. A simple example of an affective construction is the *What's X doing Y?* construction which expresses incongruity, e.g. *What's this scratch doing on the table?* (Kay & Fillmore 1999).

Other constructions, particularly focusing constructions, may contribute to the expression of affect indirectly. For example, non-defining *which*-clauses, particularly continuative ones, have been shown to encode an evaluative function in the majority of cases (Tao & McCarthy 2001). The use of such marked forms may be considered a case of *grammatical metaphor* (see below).

Grammatical metaphor

'A meaning may be realised by a selection of words that is different from that which is in some sense typical or unmarked. From this end, metaphor is variation in the expression of meanings' (Halliday 1994:341).

Halliday's concept of *grammatical metaphor*, analogous to the concept of lexical metaphor, holds that grammatical choices are made in the production of any utterance, and that such choices are meaningful. Halliday uses the term *congruent* to describe typical or unmarked forms – a congruent form can be viewed as ‘the one that is most *functionally transparent* or *motivated*’ (Veltman 2003:321). Grammatical metaphor can encode affective meaning; by employing an incongruent form which does not encode any additional ideational meaning, an affective motivation is likely to be inferred.

Semantic prosody – collostructions

The concept of collocation can be extended to constructions as *collostructions* (Stefanowitsch & Gries 2003) by considering the strength of attraction between a construction and its *collexemes* (lexis which appears in slots within the construction). Collostructional analysis shows that the concept of semantic prosody, by extension, also applies to constructions; for example, collostructional analysis of the construction *N waiting to happen* shows that it features strong negative lexical association, overwhelmingly favouring *accident* and *disaster* as collexemes (ibid.).

Features of spoken grammar

Spoken grammar differs from that of the written language and some of these differences have a bearing on the expression of affective meaning. For example, subject ellipsis, a feature of informal spoken English, frequently encodes affective meanings (Nariyama 2006). An elided utterance has a more subjective, evaluative nature (Zwicky 2005), as illustrated by the first sentence below:

> Odd that Mary never showed up.
> It is odd that Mary never showed up.

Similarly, the flexible word order of spoken English often serves evaluative functions. Carter and McCarthy (1995:151) note that tails (right-dislocated phrases) tend to occur ‘with phatic, interpersonal functions, usually in contexts of attitudes and evaluations’, for example: ‘Good winter wine *that*’.

Phonology and prosody

Phonemic modification

At the smallest phonological level, the modification of individual phonemes contributes to affective meaning. On a global level, anger (or heavily negative evaluation) increases the accuracy of articulation, while sadness reduces it (Kienast, Paeschke & Sendlmeier 1999).

Vowel duration also seems to be influenced, with happiness producing a particular lengthening effect on (stressed) vowels, followed by sadness and anger (a slight lengthening effect); conversely, fear produces a shortening effect (Kienast et al 1999).

Consonants are also modified when expressing emotions and strong attitudes. For example, a link between plosive and fricative sounds and the expression of affect, in particular aggression, has been noted (Walsh 1968) – ‘spitting out’ or ‘hissing’ words. A similar effect on the duration of voiced fricatives to that on stressed vowels has been observed, although in this case anger tends to cause a slight shortening (ibid.).

Intonation

It is notoriously difficult to establish any concrete rules regarding the use of intonation for affective purposes; although a relationship between intonation and affective meaning clearly exists, different speakers have their own ways of exploiting intonation patterns to produce affective results (Jenkins 2000). There appear to be norms at some level, however, although such norms vary from dialect to dialect (Tarone 1973).

Voice quality

The quality of a speaker’s voice – whether it is neutral, tense, breathy, whispery, harsh or creaky – is an important contributor to affective meaning. Again, the processes at work are complex, and voice quality combines with other phonological and prosodic features such as speech rate to create overall effect (Gobl & Chasaide 2003).

Other prosodic features

Marked stress, pauses and other features including those outlined above, combine to create *phonological metaphor*, which operates in a similar manner to grammatical metaphor (Veltman 2003).

Discourse-level features

Presupposition

Beyond what is directly said in a text lies a whole set of presuppositions, which together form a *presupposed world*, in which ‘the narrator has given form to an idea of what an agent and an action are, and of what an expected succession of events is’ (Marsen 2006:261). Within the presupposed world, identities are ascribed to agents by means of presupposition and relationships between agents and entities are constructed. These identities and relationships can provide the key to discovering the evaluative message of a text.

Implicature

Lexical choices (e.g. *young* versus *old*) reveal evaluative judgements; such choices are motivated, and *imply* ‘an association between these signs of identity and the actions that are ascribed to the agent’ (Marsen 2006:254). For example, an utterance such as ‘gangs of black youths were mugging elderly white women’ (Mumford & Power 2003:206) implies a connection between the identity of the agents as *black* and *youths* and their action of *mugging*.

Conversational implicature

Grice’s (1975) Co-operative Principle, with its maxims can explain much ‘unstated’ evaluation. Grice posits a set of unwritten conversation rules, or *maxims*, under the headings of *quantity*, *quality*, *relation* and *manner*. When a speaker *flouts* a maxim, the listener must deduce the reason for the speaker’s flouting of the maxim – this is a *conversational implicature*. Such conversational implicata are often attitudinal or affective.

One feature of conversational implicata is that they avoid direct expression of the speaker’s position and are therefore more difficult to challenge: ‘conversational implicata are not part of the meaning of the expressions to the employment of which they attach’ (ibid:58).

Dialogism

As speakers (or writers) are aware of their position in an ongoing dialogue, they position themselves with respect to previous statements and anticipate future responses. This process is known as dialogism (Vološinov 1986).

The degree to which speakers acknowledge the validity of differing viewpoints (*heteroglossia*) or refuse to acknowledge them at all (*monolglossia*) itself expresses affective meaning (Martin & White 2005), and increases or decreases the interpersonal cost of challenging a position (ibid.).

Context

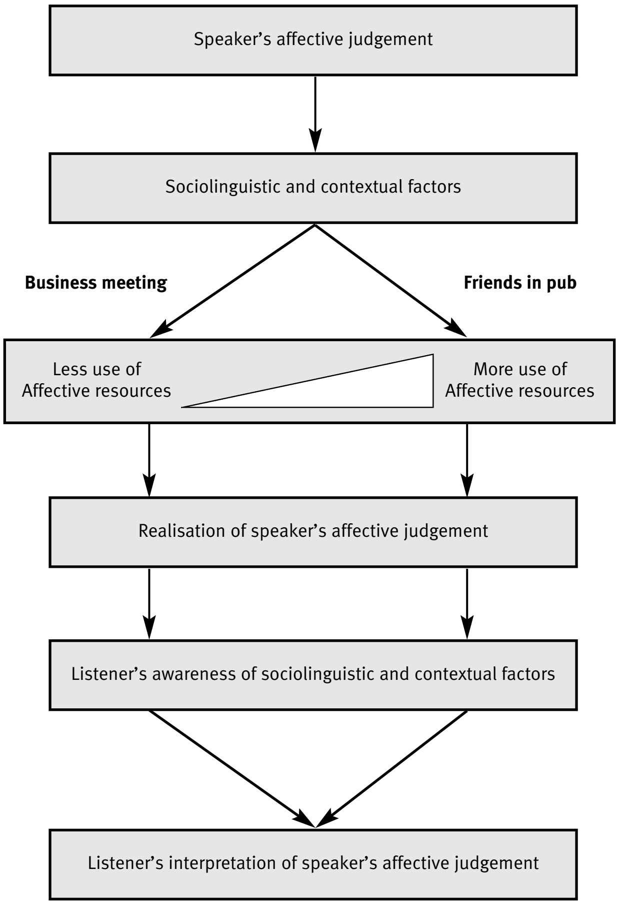
Sociolinguistic considerations

The expression of affect is not a sociolinguistic phenomenon. Sociolinguistics describes how external sociological factors influence and constrain language; affect, on the other hand, is intensely personal and *internal*. However, sociolinguistic factors constitute a key element in determining how affect is encoded, and how its expression is interpreted.

In many formal contexts, for example a business meeting, it is not considered appropriate to behave in an overtly emotional manner, so the affective resources at a speaker's disposal are circumscribed. However, this does not mean that speakers do not express attitudes; rather that the 'rules of the game' change. The result is to amplify the affective resources used – what would be considered mild in another context would be interpreted more strongly. Conversely, a group of young British men talking in a pub will often use strongly affective language without encoding a particularly strong affective meaning, and will interpret each other's utterances accordingly.

Other sociolinguistic and contextual factors – relationships in terms of familiarity, age, gender and power – also affect the nature of the expression and interpretation of affective utterances. Thus sociolinguistic and contextual factors act as 'filters' in the expression and interpretation of affective judgements, as illustrated in Figure 1.

Figure 1: Sociolinguistic and contextual filters in the expression of affect


Speaker's affective judgement
Sociolinguistic and contextual factors
Business meeting
Friends in pub
Less use of Affective resources
More use of Affective resources
Realisation of speaker's affective judgement
Listener's awareness of sociolinguistic and contextual factors
Listener's interpretation of speaker's affective judgement

'The history of a sentence'

Another important aspect of context is what Halliday (2003) described as the *history of a sentence*. A sentence can be placed in a historical context from different aspects.

Intratextual history refers to the placing of the sentence in relation to the progression of the discourse as a whole. Schematic nuances are developed, and ideational meanings previously expressed create a framework within which the sentence is interpreted.

Development history is 'the prior semiotic experience of those who enact it, as performers or receivers' (ibid:365). Development history can refer to the experience of an individual, a group or even all of humankind, and is the process by which many words and phrases develop affective connotations over time according to their usage within a particular speech community.

Other contextual features

Perhaps the most important factor in determining the type of affective resources deployed in an exchange will be the personalities of the *agents* involved. Different people express themselves differently, with more or less affective expression, or with a tendency to use more positive or negative expression than others; equally importantly, some people will adapt their utterances more according to the personality and behaviour of the other participant(s) in the exchange, or conform more to sociolinguistic norms, than others. An understanding of the nature of the participants is therefore important for a reliable analysis.

The mode of the interaction will have effects. A telephone call will require different resources from a one-to-one conversation over a cup of coffee, due to the relative availability of non-verbal resources such as gestures and facial expressions.

Methodology

The data was analysed in terms of the affective resources discussed above and how they interact to produce the affective meanings expressed in the text. The discussion presented here is summarised and narrow in scope; it does not refer to all the resources employed. For a fuller discussion, see Elliott (2008).

Context and medium

The data is taken from a BBC current affairs radio phone-in programme from 2007, featuring questions to Nick Clegg

MP, leader of the British political party the Liberal Democrats (prior to his becoming Deputy Prime Minister). The sample features a question from a female caller about immigration policy, specifically whether Clegg would 'close the borders' of the UK.

The interaction patterns within the sample are complex. While the speaker is ostensibly addressing Clegg with her question, she has another audience – the radio listeners. Indeed, it could be argued that the listeners are her primary audience, since the speaker's motivation for phoning in to such a programme seems to be to make a point rather than to make a genuine enquiry of Mr Clegg.

The medium of a radio phone-in affects the exchange. The lack of visual contact prevents the use of non-verbal communication, which means that the language itself carries all the affective meaning.

From a sociolinguistic perspective, the setting of a radio phone-in, and the position of Clegg as a senior politician, are likely to have the following effects:

- the dual audience means there are two sets of sociolinguistic norms at play – those between the speakers and the radio audience, and those between the speakers and each other
- the 'exposed' nature of the discussion, conducted in such a public forum, is likely to lead to circumspection, since the speakers will not want to appear unreasonable.

Agents

Nick Clegg has been an MEP, Liberal Democrat spokesperson for Europe (2005–06) and Home Affairs spokesperson (2006–07). In the past, he has described the issue of immigration as 'the dog-pit of British politics – a place only the political rottweilers are happy to enter' and arguing for a 'liberal managed immigration system' (Clegg 2007). The caller is Mary, a woman from Coventry. The programme was hosted by Victoria Derbyshire, a BBC Radio presenter.

Discussion

The analysis focused on the following extended turn by the caller, although the previous (and subsequent) parts of the discussion were also considered.

> 'Um … We have open borders within Europe. Millions of people can come in here potentially. Um … (unclear) I want to ask you, when did you, or any of the other two leaders, ask the people of this country if this is what they want? It's not your country. Will you close the borders within Europe if we find that we are totally swamped? Our culture and our way of life have changed beyond belief, people are scared of the extent of immigration, I believe one in four in Boston, Lincolnshire is an immigrant. Would you close our borders to people from Europe, let alone the rest of the world, if the people of this country became so distressed at … you know, I just want to know – would you close the borders, or are you so keen on Europe that you don't care how many people come here?'

The text reveals multiple objects of evaluation:

- *Immigration* and *immigrants*. Immigrants are subdivided into those from Europe and those from the rest of the world.
- *The people of this country* (or just *people*), consistently positioned as victims in the text: *people are scared of the extent of immigration: when did you … ask the people of this country if this is what they want?* The speaker positions herself with this group, which includes the listeners.
- *Politicians* – specifically *party leaders*, and in particular Clegg himself, consistently evaluated negatively.

Throughout her discourse, the speaker employs *bare assertions* – statements with no hedging employed – to create a strongly monoglossic feel, not acknowledging any alternative viewpoints. The evaluation builds through the text; we will consider two utterances of particular interest in depth (for full analysis, see Elliott 2008).

Utterance 1:

> 'Will you close the borders within Europe if we find that we are totally swamped? Our culture and our way of life have changed beyond belief, people are scared of the extent of immigration …'

The speaker uses the strongly negative term *swamped*. The term *swamped* has an interesting developmental history. It has a particular resonance in British political discourse on immigration – Margaret Thatcher was accused of racism when she used the term in 1979, and further controversy was caused in 2002 by the then Home Secretary David Blunkett's use of the term. The term *swamped* is so loaded as to create a qualitatively different feel to the discourse in affective terms. Also, *beyond belief* serves a similarly strong role.

Utterance 2:

> 'Would you close our borders to people from Europe, let alone the rest of the world, if the people of this country became so distressed at …'

Use of the *let alone* construction posits a scalar relationship between *Europe* and *the rest of the world* (Fillmore, Kay & O'Connor 1988), which would naturally be interpreted in terms of the relative desirability of immigration from the two parts of the world; this scalar relationship is reinforced by marked stress and intonation accorded to both *let alone* and *rest*.

Here, *so* is heavily marked, with marked stress, a markedly low fall, heavy sibilance on the vowel /s/ and an elongated diphthong /əu/, conveying an impression of anger (Kienast, Paeschke & Sendlmeier 1999, Walsh 1968).

The utterance is left unfinished, which naturally raises the question of how it would finish; grammatically, completion with a *that*-clause to create a cause-and-effect relationship is suggested. We can only speculate as to what the unexpressed effect would be, but we can note the following:

- The cause *if the people … became so distressed at …* evokes a fairly extreme set of circumstances, which naturalises an expectation that the response would be proportionally strong.
- The impression of an extreme response from the British people is reinforced by the fact that the utterance remains unfinished. After producing some strong, direct statements, the speaker feels unable to articulate these

consequences. She then appears to backtrack – *you know, I just want to know …* – suggesting a reasonable position on the part of the speaker, especially with the use of *just* (with a low intonational fall).

We cannot know how the speaker intended to complete the utterance, but what is important is the interpretation that the unfinished utterance, in conjunction with previous utterances, naturalises – the *perceived* attitude. This seems to be that the consequences of the people of Britain becoming so *distressed* are rather dark – too dark to be spelled out on a radio programme.

As can be seen, the utterances need to be considered in the light of the full text, plus surrounding turns and the wider context, to realise how the interaction of the different affective resources creates the full evaluative effect.

Global overview

- The use of noun phrases (*the people of this country*, *people*) and pronouns (*we*, *our*) throughout to position the people of Britain as victims of both immigration and the politicians Mary holds responsible. The use of the noun phrase *the people of this country* is interesting; concordance analysis shows that it almost exclusively occurs in political rhetoric, and that it carries a strong positive semantic prosody (Elliott 2008).
- The repeated use of bare assertions (often in conjunction with subjective statements) lends a monoglossic feel to the whole turn: the speaker does not acknowledge alternatives. This is reinforced by (phonologically) prosodic features such as a rapid speech rate for such utterances and low final falls in intonation.
- The evaluation builds throughout the turn, reaching a peak with the unfinished utterance, as the layers of evaluation interact to reinforce each other and amplify the effect.
- The complex interaction patterns and multiple audiences have an effect on the speaker as she attempts to tailor her message to the different audiences and conform to different sociolinguistic norms simultaneously (it may have been an inability to reconcile these with the intended message that led the speaker to abort the utterance).

What is particularly striking is how different affective resources interact to produce the overall effect, and how the evaluation is dependent on previous utterances (and previous texts, as in the case of *swamped*). An analysis focusing on only one or two of these areas, or on individual utterances in isolation, would not be able to account fully for the extremely strong affective meaning expressed throughout.

Conclusions

We have seen that different elements of language combine to create affective meaning in a highly interrelated manner, but that some individual elements can create a particularly strong effect which reverberates throughout the whole text. Even what is not said often can contribute greatly to the overall effect, as the unfinished utterance exemplifies.

The text we examined was a telephone-based exchange with a whole host of other contextual and sociolinguistic factors in play relating to *participants*, *medium*, (*multiple*) *audiences* and *interaction patterns*. The last two points in particular raise interesting questions for future research regarding their effects, since they apply whenever more than two people are involved in an exchange, even in a passive listening role.

These reflections raise questions regarding models of language and language proficiency – affective meaning, a central plank of communication, and often its main motivation, is underrepresented in current models and is a prime candidate for in-depth exploration, which would enrich our understanding of language as a whole. Similarly, the study of its progression as a key part of language proficiency could reap dividends, with consequences for language assessment – although obstacles such as the high context-sensitivity and deeply personal nature of affective communication are by no means easy to overcome within an assessment context.

References and further reading

Carter, R and McCarthy, M (1995) Grammar and the spoken language, *Applied Linguistics* 16 (2), 141–15.

Channell, J (2000) Corpus-based analysis of evaluative lexis, in Hunston S and Thompson, G (Eds) *Evaluation in Text: Authorial Stance and the Construction of Discourse*, Oxford: Oxford University Press, 38–55.

Clegg, N (2007) *Immigration in the 20th Century*, speech at Liberal Democrats Conference 2007, retrieved from http://www.nickclegg.org.uk/index.php?option=com_content&task=view&id=219&Itemid=45.

Coates, J (1990) Modal meaning: the semantic-pragmatic interface, *Journal of Semantics* 7, 53–63.

Dossena, M and Jucker, A (2007) Introduction, *Textus* XX, 7–16.

Elliott, M (2008) *The Expression of Affect in Spoken English: a case study*, unpublished MA thesis, King's College London.

Fillmore, C, Kay, P and O'Connor, M (1988) Regularity and idiomaticity in grammatical constructions: the case of let alone, *Language* 64 (3), 501–538.

Gobl, C and Chasaide, A (2003) The role of voice quality in communicating emotion, mood and attitude, *Speech Communication* 40, 189–212.

Goldberg, A (2003) Constructions: a new theoretical approach to language, *Trends in Cognitive Sciences* 7 (5), 219–224.

Grice, H (1975) Logic and Conversation, in Cole, P and Morgan, J (Eds) *Syntax and Semantics Volume 3: Speech Acts*, London: Academic Press, 41–58.

Halliday, M (1978) *Language as a Social Semiotic*, London: Arnold.

Halliday, M (1994) *An Introduction to Functional Grammar* (2nd ed.), London: Arnold.

Halliday, M (2003) *On Language and Linguistics* (edited by Webster, J), London: Continuum.

Hunston, S (2007) Using a corpus to investigate stance quantatively and qualitatively, in Englebretson, R (Ed.) *Stancetaking in Discourse*, Amsterdam: John Benjamins, 27–48.

Jenkins, J (2000) *The Phonology of English as an International Language*, Oxford: Oxford University Press.

Kay, P and Fillmore, C (1999) Grammatical constructions and linguistic generalizations: the what's X doing Y? construction, *Language* 75 (1), 1–33.

Kienast, M, Paeschke, A and Sendlmeier, W (1999) Articulatory reduction in emotional speech, *EUROSPEECH '99*, 117–120.

Larsen-Freeman, D and Cameron, L (2008) *Complex Systems and Applied Linguistics*, Oxford: Oxford University Press.

Marsen, S (2006) How to mean without saying: presupposition and implication revisited, *Semiotica* 160, 243–263.

Martin, J and White, P (2005) *The Language of Evaluation: Appraisal in English*, Basingstoke: Palgrave Macmillan.

Mumford, K and Power, A (2003) *East Enders*, Bristol: The Policy Press.

Nariyama, S (2006) Pragmatic information extraction from subject ellipsis in informal English, *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, 1–8.

Perez, A (1990) Time in motion: grammaticalisation of the be going to construction in English, *La Trobe University Working Papers in Linguistics* 3, 49–64.

Stefanowitsch, A and Gries, S (2003) Collostructions: investigating the interaction of words and constructions, *International Journal of Corpus Linguistics* 8 (2), 209–243.

Tao, H and McCarthy, M (2001) Understanding non-restrictive which-clauses in spoken English, which is not an easy thing, *Language Sciences* 23, 651–677.

Tarone, E (1973) Aspects of intonation in Black English, *American Speech* 48 (1–2), 29–36.

Thompson, G and Hunston, S (2000) Evaluation: An introduction, in Hunston, S and Thompson, G (Eds) *Evaluation in Text: Authorial Stance and the Construction of Discourse*, Oxford: Oxford University Press, 1–27.

Veltman, R (2003) Phonological metaphor, in Simon-Vandenbergen, A-M, Taverniers, M and Ravelli, L (Eds) *Grammatical Metaphor*, Amsterdam: John Benjamin, 311–335.

Vološinov, V (1986) [1929] *Marxism and the Philosophy of Language*, Cambridge, MA: Harvard University Press.

Walsh, M (1968) Explosives and spirants: primitive sounds in cathected words, *Psychoanalytic Quarterly* 37, 199–211.

Wierzbicka, A (1987) Boys will be boys: 'radical semantics' vs. 'radical pragmatics' *Language* 63 (1), 95–114.

Zwicky, A (2005) Saying more with less, *Language Log*, retrieved from http://158.130.17.5/~myl/languagelog/archives/2005_03.html.

Peer–peer interaction in a paired Speaking test: The case of FCE

EVELINA D GALACZI RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

This short summary is based on a doctoral thesis submitted to Columbia University, New York City (US) in 2004. The PhD was supervised by Professor James Purpura.

This discourse-based study, which was undertaken as part of a doctoral degree, investigated paired test taker discourse in the First Certificate in English (FCE) Speaking test. Its primary aim was to focus on fundamental conversation management concepts, such as overall structural organisation, turn-taking, sequencing, and topic organisation of the paired test taker interaction. The analysis highlighted global patterns of interaction in the peer test taker dyads and salient discourse features of interaction. The three distinct patterns of interaction which emerged were termed 'collaborative', 'parallel', and 'asymmetric'. The patterns of interaction were distinguished based on the dimensions of mutuality and equality, and were conceptualised as continua ranging from high to low. In addition, the dimension of conversational dominance, operationalised as 'participatory', 'sequential', and 'quantitative', was found to intersect with the dimensions of mutuality and equality, leading to sub-groups within each interactional pattern of high or low conversational dominance. The second goal of the study was to investigate a possible relationship between the patterns of peer–peer interaction and the FCE score for 'interactive communication' (IC). The aim was to understand more accurately the relationship between the discourse generated by the task and the scores for 'interactive communication', and to provide some validity evidence for the IC scores. The results showed that the high-scorers mostly oriented to a collaborative pattern of interaction, while the low scorers generally oriented to a parallel pattern of interaction, as would have been expected. The significance of the study lies in the deeper understanding it provides of paired oral test interaction in the FCE and the construct of conversation management. This study also holds implications for FCE examiner training as it provides insights which could lead to more accurate and consistent assessment of FCE candidate output. A further contribution of the present study is the recommendations it provides for the performance descriptors used for 'interactive communication' in the FCE assessment scales, which would ultimately lead to a fairer test. For more details on this study, see Galaczi (2003, 2008).

References

Galaczi, E D (2003) Interaction in a paired speaking test: the case of the First Certificate in English, *Research Notes* 14, 19–23.

Galaczi, E D (2008) Peer–Peer Interaction in a Speaking Test: The Case of the First Certificate in English Examination, *Language Assessment Quarterly* 5 (2), 89–119.

Second language acquisition of dynamic spatial relations

IVANA VIDAKOVIĆ RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

This short summary is based on a doctoral thesis submitted to the University of Cambridge (UK) in 2006. The PhD was supervised by Dr Henriëtte Hendriks.

The aim of this thesis is to shed light on the nature of adult second language acquisition, factors guiding the acquisition and the ways in which these factors interact. This is achieved through exploring how English learners of Serbian and Serbian learners of English acquire another way of expressing dynamic spatial relations (motion) in a second (foreign) language.

Talmy (1985) divides languages into:

a. satellite-framed, typically encoding Path in satellites and Manner in motion verbs (e.g. *The bottle floated out*) and

b. verb-framed, typically encoding Path in motion verbs and Manner, if expressed at all, outside the verb (e.g. *La botella salió flotando – The bottle exited floating*).

English and Serbian were both classified as satellite-framed languages within Talmy's typology. However, recent research revealed that Serbian differs to a certain extent from English as to where Manner and Path are typically expressed (Filipović 2002), and as to the frequency of expression of Manner. Therefore, Filipović (2002) reclassified Serbian placing it midway in the continuum satellite-framed›Serbian›verb-framed. The contribution of the non-acquisition part of the thesis resides in providing further support for the reclassification of Serbian, based on the analysis of the spoken mode of language use and systematic examination of attention to Manner (as reflected in the frequency of Manner mention). The findings show that:

a. when they want to express Manner in boundary-crossing situations (e.g. entering, exiting, crossing), Serbian native speakers most frequently opt for the verb-framed pattern of expressing Path in the verb and Manner outside it when using their mother tongue, and

b. they omit Manner information considerably more frequently than English native speakers when speaking in their mother tongue, even when Manner is not inferable from the context.

Using the Interlanguage approach, the main, acquisition-related part of the thesis examines how lower-intermediate, upper-intermediate and advanced learners express motion at a given stage of the acquisition process, how their linguistic means develop and what factors influence the acquisition. According to this approach, which has proved fruitful for analysing the acquisition process of beginners, learners' interlanguage and its development over time are systematic. This systematicity cannot be directly related to either the first or the second language. The acquisition paths exhibit similarities across different (first – L1, and second – L2) language pairings, being influenced mostly by universal, and only marginally by language-specific factors, since the interlanguage of beginners is syntactically and semantically a very simple system. Previous studies on higher-level learners, whose interlanguages are more complex syntactically and semantically, document mostly language-specific influences. The present thesis set out to investigate whether universal characteristics of learners' development persist among learners beyond the beginning stage, or whether only language-specific influences hold sway, how all of them manifest and what their scope is. Since the learners examined are beyond the beginning stage, the over-arching hypothesis was that language-specific influences would be stronger than among beginners and acquisition paths not so homogenous, yet factors other than first or second language may bring out similarities in the interlanguages and acquisition paths of learners with different first and second languages.

One of the contributions of the present thesis resides in showing that even the interlanguage of learners beyond the beginning stage shows similarities unrelated to the first or second language, and also that it exhibits a rich interplay of both language-specific (L1/L2) and universal factors. For example, both English and Serbian learners mostly prefer the satellite-framed, English pattern (e.g. *run into X*) to the verb-framed pattern favoured by Serbian native speakers when using their L1 (e.g. *go running into X*). In this way, learners resort to the economy-of-form strategy[1] opting for a pattern that is more economical by being shorter, syntactically simpler and thus easier for processing (production/understanding). It is in the domain of linguistic attention to Manner that a language-specific influence (L1 influence) is at its strongest at times, being clearly visible even among the advanced English and Serbian learners. In addition, the findings reveal that L2 learners undergo not only linguistic reorganisation, but also a change in the degree of linguistic attention to Manner (increasing/decreasing frequency of Manner mention) with increasing proficiency levels.

Besides theoretical implications for the field of second language acquisition, this thesis has also practical implications for teaching the linguistic devices expressing dynamic spatial relations in the two languages. For more details on this study see Filipović & Vidaković (2010).

[1] This term was first used in Vidaković (2006).

References

Filipović, L (2002) *Verbs in motion expressions: structural perspectives*, unpublished PhD dissertation, University of Cambridge.

Filipović, L and Vidaković, I (2010) Typology in the L2 classroom: Second language acquisition from a typological perspective, in Pütz, M and Sicola, L (Eds) *Cognitive Processing in Second Language Acquisition*, Amsterdam/Philadelphia: John Benjamins, 269–293.

Talmy, L (1985) Lexicalization patterns: Semantic structure in lexical forms, in Shopen, T (Ed.) *Language typology and syntactic description: Grammatical categories and the lexicon*, Cambridge: Cambridge University Press, 57–149.

Vidaković, I (2006) *Second Language Acquisition of Dynamic Spatial Relations*, unpublished PhD dissertation, University of Cambridge.

Demonstrating cognitive validity of IELTS Academic Writing Task 1

GRAEME BRIDGES ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

Introduction

This paper is based on a Master's thesis submitted to Anglia Ruskin University, Cambridge, UK, in October 2008. The research was funded by Cambridge ESOL. The MA was supervised by Dr Sebastian Rasinger.

This study further examines the validity of IELTS Academic Writing Task 1, the first of two compulsory tasks that are designed to test the writing ability of those wishing to study or work in the medium of English. The study makes use of Weir's (2005) socio-cognitive validity framework and focuses on cognitive validity by investigating the appropriateness of the cognitive processes required to complete IELTS Academic Writing Task 1. As a secondary research goal, the processes required to address two different kinds of visual input employed in Task 1 – a graph and a diagram – are compared.

The study uses two research instruments – a verbal protocol technique and a questionnaire which together provide qualitative and quantitative data.

The findings demonstrate that Task 1 does engage those cognitive processes that are deemed essential in the target language use domain. The study reveals that this task is essentially a knowledge telling exercise, so some processes, especially organising, are under-represented with this task type. It also shows that there seem to be some differences in IELTS candidate perception regarding the data and diagram task type although such differences are statistically insignificant.

Although a relatively small-scale research project, the study not only provides further evidence of cognitive validity of this task type but also raises questions for further research.

Literature review

The socio-cognitive approach

Cambridge ESOL has for almost the last 20 years used the VRIP (Validity, Reliability, Impact and Practicality) approach to validating its tests (Saville 2003) with validity 'generally considered to be the most important quality' (ibid:65). In this framework, validity, although seen as a unitary concept, is configured on the basis of three types of evidence:

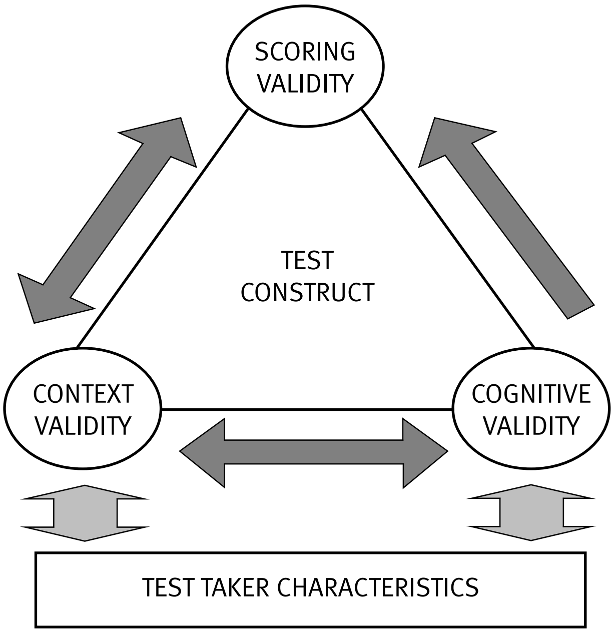
- **construct:** the extent to which the test scores reflect the test takers' underlying language knowledge and abilities based on a model of communicative language ability (see Bachman 1990)
- **content:** the extent to which the content of the test represents the target language use domain
- **criterion:** the extent to which the test scores are correlated with an external criterion that measures the same knowledge and abilities.

Weir's approach reconfigures construct validity along three dimensions – context, cognitive processing and scoring – and shows how they interact with each other thereby demonstrating the unitary nature of validity (Weir 2005, Weir, O'Sullivan, Jin & Bax 2007). In this model, the construct does not just reflect the underlying traits of communicative language ability but is the result of trait, context and score. The 'trait-based' or 'ability' approach to assessment is thus reconciled with the 'task-based' or 'performance' approach. An interactionalist position (Chapelle 1998:43) is thus adopted whereby the construct resides in the interaction between the underlying cognitive ability and the context of use – hence the socio-cognitive model (see Shaw & Weir 2007:2).

Figure 1 depicts how the components that make up construct validity join together both temporally and conceptually. The arrows indicate the relationship between the components with the timeline running from bottom to top. The *test taker characteristics* box connects directly to the *cognitive* and *context validity* boxes because 'these individual characteristics will directly impact on the way the individuals process the test task set up in the *context validity* box' (Weir 2005:51).

In this framework cognitive validity involves collecting both *a priori* evidence on the mental processing activated by the test and *a posteriori* evidence involving statistical analysis of scores following test administration. As this study concentrates on just considering *a priori* evidence, score analysis does not form part of the methodology.

Figure 1: Construct validity components



(Shaw & Khalifa 2007)

The elements making up construct validity can be seen to be symbiotically related in that decisions taken in terms of task context will impact on the processing that takes place in task completion. Likewise scoring criteria where known to the test taker will impinge on cognitive processing. Taken together 'the more evidence collected on each of the components of this framework, the more secure we can be in our claims for the validity of a test' (Weir 2005:47).

Models of second language writing

Before the 1960s writing was often conceptualised as transcribed speech and was viewed as 'decontextualised' (Ellis 1994:188) and product-oriented with final texts seen as 'autonomous objects' where various elements were organised according to a 'system of rules' (Hyland 2002:6). Writing is now seen as essentially a communicative act. A written text is therefore viewed as discourse in that the writer attempts to engage the reader using linguistic patterns influenced by a variety of social constraints and choices (writer's goals, relationship with audience, content knowledge, etc.). Any model of writing needs to account for these contextual factors and see writing as a social act.

A model of writing also needs to take account of the internal processing writers undertake. A recent model from Field (2004) is based upon information processing principles from psycholinguistic theory. He provides a detailed account of the stages a writer proceeds through:

- **macro-planning:** ideas gathering and identifying major constraints (genre, readership, goals)
- **organisation:** ordering ideas and identifying relationships between them
- **micro-planning:** focusing on the part of the text (paragraph and sentence) about to be produced
- **translation:** converting prepositional content held in abstract form to linguistic form
- **monitoring:** checking mechanical accuracy and overall coherence
- **revising:** adjusting text as a result of monitoring.

These stages of executive processing are the basis of Shaw and Weir's (2007) conceptualisation of the cognitive validity component of the socio-cognitive framework for writing and as such inform the methodology outlined below.

A parallel strand of research focuses not on the stages of the writing process *per se* but how these relate to different levels of language proficiency. Eysenck & Keane (2005:418) argue that it is the planning process that differentiates the skilled from the unskilled writer. Scardamalia & Bereiter (1987) describe two major strategies, knowledge telling and knowledge transforming, which occur mainly at the planning stage and help to identify the processing of skilled writers and the less able. In knowledge telling the writer plans very little and is concerned mainly with generating content from remembered existing resources in terms of content, task and genre. In knowledge transforming the skilled writer considers the complexities of a task as well as content, audience, register and other relevant factors in written communication.

IELTS-related writing research

As a high-stakes test IELTS has always attracted attention from researchers including those who have focused just on the writing component. Much of the research has been generated by the IELTS partners themselves thus demonstrating their commitment to the continual improvement of the test (see for example Taylor & Falvey (2007) for a collection of IDP and British Council joint-funded research reports on IELTS Writing). In 2005, the assessment criteria and rating scales were revised in IELTS Writing largely as a consequence of these and other research findings. Many of the inevitable criticisms that a high-stakes test such as IELTS attracts were addressed in 2005 but some issues concerning cognitive validity still remain.

Of the two tasks in IELTS Academic Writing most research has been conducted on Task 2, the short essay. Being the longer of the two in terms of time allocation (40 minutes) and word length (250 words) it generates a greater sample of L2 writing. There have therefore been several *a posteriori* studies on Task 2 candidate scripts (see Mayor, Hewings, North, Swann & Coffin 2006). Task 2 also carries the heavier weighting in scoring, one of the justifications for Moore & Morton's (2006) *a priori* study on test task authenticity. Weir et al (2007) were the first to use a specially designed cognitive validity-based questionnaire in their study of comparability of word-processed and pen & paper IELTS writing. In that study, they compared candidate scores on two Task 2 prompts (*a posteriori*) as well as a quantitative and qualitative analysis of the questionnaire responses (*a priori*). This questionnaire forms the basis of one of the research instruments used in my study.

Task 1 on the other hand has generated relatively less research interest and apart from some internal Cambridge ESOL validation studies, it has always been researched alongside Task 2. Of greatest relevance to the present study is Mickan, Slater and Gibson's (2000) *a priori* study examining the readability of test prompts (Task 1 and 2) and test-taking behaviours of intending IELTS candidates using verbal protocol analysis. This study essentially focused on the context validity parameters of task input emphasising the 'socio-cultural influences on candidates' demonstration of their writing ability' (ibid:29). As many aspects of IELTS Writing have evolved since this study, including the rubric, it would be interesting to see how candidates perceive Task 1 now.

Methodology outline

From the above literature review I have located a research area where much has already been explored. However, at the time of writing the thesis, cognitive validity of IELTS Task 1 had not been investigated before, to my knowledge, and there had not been an attempt to apply both qualitative and to some extent quantitative methodologies in one study to generate *a priori* evidence supporting the cognitive validity of just the Task 1 in IELTS Academic Writing.

Two research instruments were employed (see Table 1 below). Firstly, verbal protocol analysis (VPA) was utilised with four IELTS preparation students as they wrote a response to one of two Academic Writing prompts: one data prompt represented as a graph and one diagrammatic prompt (see Figures 2 and 3). The aim was to provide rich qualitative data on the cognitive processes undertaken by IELTS candidates. The same four students then responded to a questionnaire which sought to further elicit their thought processes. By distributing this questionnaire to 56 other students, quantitative as well as qualitative data was generated.

Table 1: Data collection methods

Methodology	Instrument	Participants
Qualitative	'Think aloud' verbal protocols (concurrent/ non-mediated)	4 IELTS candidates at various levels of proficiency and L1 background
Quantitative/ Qualitative	Cognitive-processing questionnaire	4 candidates above + 56 other candidates of varying levels and L1 background

'Think aloud' verbal protocols

Verbal protocol analysis is an introspective technique that is well-suited to obtain evidence of cognitive processing as part of construct validation. A participant is asked to 'talk aloud' or 'think aloud' as they carry out a task with their utterances comprising the 'protocol'. 'Verbal protocol' is the data gathered under these conditions. These verbalisations can be seen as an accurate record of the participant's thought processes. It is important to stress, however, that 'individuals cannot report their own cognitive processes' and that it is for the researcher to 'infer cognitive processes and attended information' (Green 1998:4). In other words participants are required to verbalise their thoughts and not the processes leading to those thoughts.

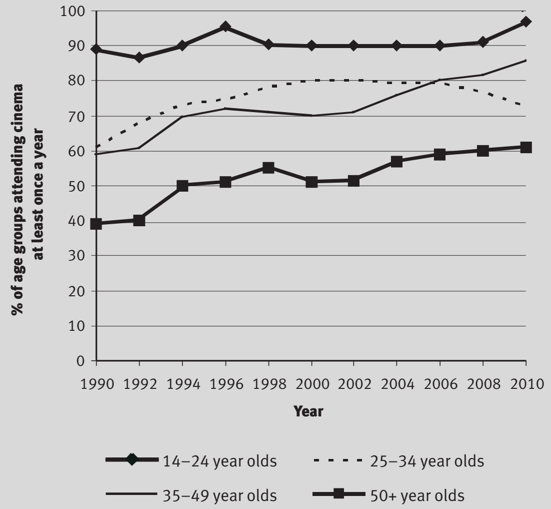
Figure 2: Data input task – cinema attendance

You should spend about 20 minutes on this task.

The graph below gives information about cinema attendance in Australia between 1990 and the present, with projections to 2010.

Summarise the information by selecting and reporting the main features, and make comparisons where relevant.

Write at least 150 words.



(Source: *IELTS Scores Explained* 2006)

Figure 3: Diagrammatic input task – brick manufacturing

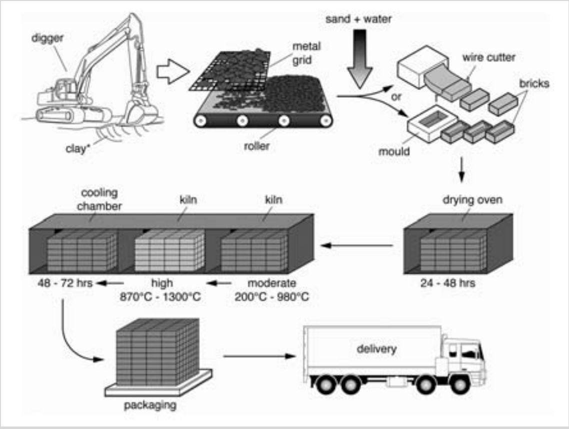
You should spend about 20 minutes on this task.

The diagram below shows the process by which bricks are manufactured for the building industry.

Summarise the information by selecting and reporting the main features, and make comparisons where relevant.

Write at least 150 words.

Brick manufacturing



*Clay: type of sticky earth that is used for making bricks, pots, etc.

(Source: *IELTS Scores Explained* 2006)

In the study the participants were asked to verbalise their thoughts concurrently as they wrote their responses to one of two non-live IELTS Academic Writing Task 1s. Their verbalisations were audio-recorded generating a set of protocols making up a body of qualitative data. Concurrent reports are generally regarded as more reliable than retrospective reports in that data is not reliant on recovering thoughts from memory. These reports were supplemented by field notes (e.g. instances of underlining, crossing out and insertions were recorded). As it was important not to interfere with the thought processes that were being explored, a non-intrusive approach was used (i.e. 'non-mediated') where prompting only occurred during long pauses and included the request to 'keep talking' or occasionally 'speak louder'. A quick debriefing at the end of the recording session took place where subjects were asked to comment on the task and the research procedure. After this, the participants were asked to complete a questionnaire.

After the data was collected, the recordings were transcribed and data was segmented according to their correspondence to single thought processes. The unit of analysis for segmentation was sometimes a word, phrase, clause, sentence or even 2–3 sentences. Each segment was delineated with a '/' and timed. In order to facilitate analysis, a coding scheme was developed by focusing on each protocol at a time and attempting to describe each segment as a thought process.

This involved four iterations of re-coding until a scheme was established that accounted for all four sets of protocols. Green (1998:70) emphasises that it is important at this stage to keep 'any theoretical assumptions to a minimum' as otherwise there is the danger of ignoring those verbalisations that are inconsistent with a particular hypothesis.

The coding that finally emerged consisted of each protocol being divided into three phases – pre-writing, writing and post-writing – and was labelled PreW, W and PostW respectively. Each thought process was then assigned a number so that PreW1 for example referred to the process of 'Reading (part of) the introductory background to the visual input'. As well as code labels and length of time, comments from the field notes were also collated. For example, the beginning of a participant's protocol was presented as follows:

Segment	Time	Verbal protocol	Code	Length of time	Comments
001	00.00	OK. Writing Task 1. You should spend about 20 minutes on this task/	PreW3	00.08	
002	00.08	The diagram below shows the process by which bricks are manufactured for the building industry/	PreW1	00.10	Underlines 'bricks' on task
003	00.18	Summarise the information by selecting and reporting the main features, and make comparisons where relevant/	PreW3	00.17	Underlines 'make comparisons' and circles 'main features' on task

Two participants (1 and 3) thought aloud as they wrote a response to the data input task (Figure 2) and two participants (2 and 4) responded to the diagrammatic Academic Writing Task 1 (Figure 3).

All were IELTS preparation students at Anglia Ruskin University in Cambridge. More demographic information on the participants is provided in Table 2 below.

Table 2: Demographic data of participants

Participant	Nationality/ Language	Gender	Age	Reasons for taking IELTS
1	Swiss French	F	20	Hoping to do BA in Business Management in London
2	Mongolian	F	29	Hoping to do an MA in Modern Society and Global Transformation at Cambridge
3	Korean	M	23	Wants to do BA in Sports Management at Loughborough
4	French	M	20	Wants to improve English while in UK for a year

Cognitive processing questionnaire

For this part of the study, I adapted the 38-item cognitive processing questionnaire (CPQ) designed by Weir et al (2007:321). The questions are grouped to reflect the cognitive processes that writers are hypothesised to undergo and are identified in the CPQ as one of Field's (2004:329) six stages outlined previously. For example, question 21 (see below) is one of several that focuses on the translation phase:

> I felt it was easy to express ideas using the correct sentences.
> 1. Strongly disagree 2. Disagree 3. No view 4. Agree 5. Strongly agree

Each stage is represented by at least four questions in order to enhance the reliability of the questionnaire, as a single question is always susceptible to bias.

A further advantage of this procedure is its uni-dimensionality in that all the questions measure in the same direction. Each item can therefore be scored from 1 to 5 (except Question 12 which elicited a yes/no response). The higher the score, the more favourable is the attitude. This in turn means that a frequency count can be carried out for the number and percentage of respondents who choose each option of each question. The mean value of responses to each question can then be calculated to reveal the tendency of the responses with the proviso that a minimum number of 30 respondents are sourced.

For those four who participated in the think aloud procedure, this questionnaire was administered afterwards in order to avoid the possible contamination of the protocols. As well as to these four participants, I distributed this questionnaire to several language schools that run IELTS preparation courses in order to generate some quantitative data.

A total of 60 IELTS preparation students of varied nationalities studying in the UK (44 students) and Hong Kong (16 students) wrote a response to either the data input or diagrammatic writing tasks (see Table 3). They then completed one of two questionnaires depending on the task they had responded to.

Table 3: Breakdown of respondents by language institute and task type

IELTS preparation course provider	No of respondents		
	Data	Diagram	Total
Eurocentres, Cambridge	4	4	**8**
Anglia Ruskin University, Cambridge	8	10	**18**
St Giles, central London	9	9	**18**
Centre for Language in Education, Hong Kong Institute of Education	8	8	**16**
Total	**29**	**31**	**60**

Data collection and analysis

'Think aloud' verbal protocols

From each of the protocols collected, the instances where a coding category was applied were ranked in order of time spent. This was supplemented with data on the frequency of instances so that together these rather crude measures could provide some indication of the prevalence of certain thought processes. This information was collated for each writing phase for each participant.

For the purposes of exemplification, the findings of each writing phase based on the verbal protocol of Participant 1 are summarised in Table 4a, Table 4b and Table 4c. Of the 23.09 minutes she took to complete the task, she spent 03.41 minutes planning her response (see Table 4a). There is evidence of macro-planning in that she clarified the task requirements by reading the task-specific rubric (PreW1 and PreW2) and the graphical input (PreW5). She attempted to interpret the data (PreW8) and summarise it (PreW9). This was the only protocol where there was evidence of topic definition (PreW10) where the writer generates ideas by utilising world knowledge. However, at no time did she write any notes although she did claim in the debriefing that she made notes in her head.

Just over 75% of the time (17.33 minutes) was spent actually writing (see Table 4b), of which she spent 07.54 minutes engaging in translating – the actual conversion of abstract ideas to linguistic form (W2 22 instances). The second most common thought process was rehearsing a linguistic form before writing. There were eight instances of this (W1) which generally occurred before the actual putting of pen to paper. There were however some overt examples of micro-planning where the writer broke off mid-sentence, tried to find a phrase to continue the sentence, went back to the task and read the instructions and then made a goal statement, previewed an idea before finally writing. This highlights the dynamic nature of writing where the text becomes part of the context thus compelling the writer re-visit the task, the instructions, goals and their memory before they can continue encoding their thoughts.

Table 4a: Pre-writing phase (3 minutes 41 seconds/15.91% of overall time on task)

Code	Coding category	Length of time	Frequency
PreW10	Defining the topic	00.46	3
PreW5	Reading (part of) the visual input	00.36	4
PreW8	Interpreting feature(s) of visual input	00.31	5
PreW9	Summarising feature(s) of visual input	00.23	2
PreW2	Re-reading (part of) the introductory background to the visual input	00.13	1
PreW6	Re-reading (part of) the visual input	00.11	1
PreW1	Reading (part of) the introductory background to the visual input	00.10	1
PreW7	Previewing potential linguistic form(s)	00.07	1

Table 4b: Writing phase (17 minutes 33 seconds/75.81% of overall time on task)

Code	Coding category	Length of time	Frequency
W2	Converting ideas into text	07.54	22
W1	Rehearsing a linguistic form before writing	01.58	8
W11	Interpreting feature(s) of visual input	01.28	5
W14	Reviewing grammatical/lexical correctness after writing some text	01.22	4
W4	Previewing a concept before writing some text	00.54	1
W17	Making a goal statement	00.50	4
W13	Reviewing grammatical/lexical correctness while writing some text	00.37	3
W3	Attempting to retrieve a linguistic form from memory	00.25	2
W7	Reading (part of) the standard instructions	00.25	1
W15	Reviewing informational content while writing some text	00.11	1
W20	Monitoring the word count	00.05	1

As well as micro-planning there are also examples of monitoring during (W13) and after writing some text (W14). While writing there were occasions where the writer self-corrected some errors e.g. *The graph illustrate illustrates erm/* (W13). This is an example of low level monitoring involving mechanical accuracy such as punctuation, spelling and syntax. However, the monitoring that occurred after some text had been written does require more attentional resources as it involves checking cohesion between sentences and within sentences e.g. the writer in her final paragraph prepared to write 'To conclude', realised that the previous paragraph began with 'To conclude' so replaced it with 'To compare' some 3 minutes after originally beginning the penultimate paragraph.

The degree of monitoring however did not seem to extend to any consideration of the reader or to goals set earlier. Nevertheless there is evidence of an evolving orientation towards goals. There are four instances of this where the writer prompts herself: *to make a difference, write one more sentence then a conclusion, draw a comparison and put it in my conclusion* (W17).

The sheer complexity of writing is further evidenced with this participant in that she prompted herself twice to retrieve a linguistic form from her long-term memory (W3), felt the need to read the standard instructions for the first time (W7), reviewed the informational content of a piece of text (W15) and was aware of the need to monitor the word count (W20).

This participant was only one of two subjects who devoted any time to the post-writing phase (see Table 4c) although she had to prompt herself to do this (PostW1). She mostly spent the time correcting errors (PostW3) although there were a couple of instances where she read her script making no corrections (PostW2). In her debriefing she thought that her response was short and that she didn't have enough time to count the number of words.

Table 4c: Post-writing phase (1 minute 55 seconds/8.28% of overall time on task)

Code	Coding category	Length of time	Frequency
PostW3	Editing (part of) text	01.02	3
PostW1	Making a goal statement	00.18	2
PostW2	Reading (part of) text	00.17	2

Overall there is strong evidence from this and from the other three participants that all but one of the cognitive processes outlined in Weir (2005) and Field (2004) are being employed. The only process where there was very little evidence was of organising – this was also the case in Mickan et al's (2000) study which concentrated on Task 2, a longer task requiring knowledge transforming skills. Perhaps even more so for Task 1, candidates are unlikely to write notes or mentally plan an outline. What was striking from all the participants was the perception that there was not enough time so perhaps organising the response was sacrificed due to that. However, from the participants' scripts and also from some of their goal statements there was still some evidence of the provisional outlining of ideas.

The findings based on verbal reports of all four participants showed that there did not seem to be any striking dissimilarities in thought processes between those taking the data input task as opposed to the diagram. Differences were largely based on writing competence with the more skilled writers such as Participant 1 engaging more in macro-planning and monitoring than the less skilled (for more information see Bridges 2008).

Not surprisingly, the protocols collected in this study provide stronger evidence of knowledge telling than knowledge transforming. Task 1 is after all designed to facilitate the transfer of assembled information from a visual input to a verbal written output.

It must be emphasised, however, that as this study involved just four participants it should be seen as exploratory and any conclusions drawn are tentative. There are also drawbacks with the methodology of VPA itself which need to be considered in any conclusion.

Cognitive processing questionnaire

The design of the two questionnaires was aimed at investigating, through participants' self-reports, the extent of the cognitive processes they employ in responding to two types of the Academic Writing Task 1. Table 5 below summarises the different stages and the questions designed to elicit respondent behaviour.

Table 5: Stages involved in writing and questions designed to elicit candidates' behaviour

Stages	Question No.
Macro-planning	1–9
Organising	10–15
Micro-planning	16–19
Translating	20–26
Monitoring & Revising	27–38

From the frequency data collected, the percentage of agreement for each question was obtained by adding up the percentage of those expressing agreement and strong agreement. This was done by task type and as a total and is presented in Tables 6, 7, 8, 9 and 10 overleaf. Preceding each table is a summary of the data highlighting the main findings with some tentative speculation as to the reasons for the results.

Macro-planning

In the goal-setting part of this stage (questions 1–5, see Table 6) there is generally quite high to very high agreement among the respondents. It does seem that many of these preparation students do read the instructions very carefully and attempt to interpret both these and the visual input so that they can meet the task requirements. This seems to be especially true of those who responded to the diagrammatic input.

A very low proportion of candidates seem to utilise world knowledge or consider the genre constraints when responding to Academic Writing Task 1s. Regarding the question of topic knowledge (Q6), it could be argued that low levels of agreement are actually a good thing as IELTS Writing tasks should not be seen to favour candidates from any specific discipline. Tasks have to be about something but not at a level where specialised knowledge would create bias.

Of more concern perhaps is the low level of knowledge about this task type which is a 150-word descriptive summary (cf. question 8 in Table 6 overleaf). Interestingly, more candidates, albeit very marginally, seemed to be more familiar with the diagrammatic task type than the data input.

Organising

A not particularly clear picture emerges from this sample during this organising stage (see Table 7). For questions 10 and 11, which elicit information on whether the writer starts to generate their ideas after the macro-planning phase above, it seems that about a third of the students report that they engage in these activities.

Questions 12 and 13 reveal that just over half do plan an outline either on paper or as mental notes and that just over 50% have thought of their ideas before they plan their outline. These ideas may well be incomplete (see question 10) or not well-organised (question 11) but there does seem to be some provisional organisation of ideas.

Not surprisingly, as 51.7% reported that they thought of most of their ideas before planning an outline, only 29% mostly thought of ideas while planning an outline. An equally low percentage thought of their ideas in English. This is not altogether surprising. L2 writers, especially unskilled ones, may experience a heavy cognitive load in simply encoding their thoughts as they write so are unlikely to plan for writing in English.

Micro-planning

This level of planning takes place as the text evolves at both the paragraph and sentence level while also taking into account decisions made in macro-planning. Perhaps the most interesting finding is the substantial difference in responses between those who responded to the diagrammatic task, of whom 58.1% thought it was easy to put their ideas in good order, and the data task respondents, of whom only 17.2% thought it was easy (see question 19 in Table 8). It could be surmised that the diagrammatic task does offer more scaffolding than the data task although interestingly more data task respondents reported being able to put their ideas or content in good order (46.7% to 31.1%, see question 17) but that of course does not necessarily mean it was easy.

Table 6: Macro-planning (Agreement with questions 1–9)

Question		agree or strongly agree		
		Data (n=29)	Diagram (n=31)	Total (n=60)
1	I FIRST read the instructions very slowly considering the significance of each word in it.	55.2%	77.5%	66.7%
2	I thought of WHAT I was required to write after reading the instructions and visual input.	79.3%	80.6%	80.0%
3	I thought of HOW to write my response so that it would respond well to the instructions.	79.3%	71.0%	75.0%
4	I thought of HOW to satisfy readers or examiners.	55.2%	42.0%	48.3%
5	I was able to understand the instructions for this writing test completely.	69.0%	80.7%	75.0%
6	I know A LOT about this topic, i.e., I have enough ideas to write about this topic.	24.1%	16.1%	20.0%
7	I felt it was easy to produce enough ideas for the Task 1 from memory.	17.2%	35.5%	26.6%
8	I know A LOT about this task type, i.e. I know how to write a descriptive summary of data (chart, diagram, table)/diagrams (process, map, plan).	24.1%	25.8%	25.0%
9	I know A LOT about other types of IELTS Academic Writing Task 1s e.g., diagrams (process, map, plan)/data (chart, diagram, table).	27.5%	32.2%	30.0%

Table 7: Organising (Agreement with questions 10–15)

Question		agree or strongly agree		
		Data (n=29 for Q10–11) (n=15 for Q12–15)	Diagram (n=31 for Q10–11) (n=16 for Q12–15)	Total (n=60 for Q10–11) (n=31 for Q12–15)
10	Ideas occurring to me at the beginning tended to be COMPLETE.	34.5%	32.2%	33.4%
11	Ideas occurring to me at the beginning were well ORGANISED.	31.0%	45.1%	38.3%
12	I planned an outline on paper or in my head BEFORE starting to write.*	51.8%	51.6%	51.7%
13	I thought of most of my ideas for the task BEFORE planning an outline.	60.0%	43.8%	51.7%
14	I thought of most of my ideas for the task WHILE I planned an outline.	33.3%	25.1%	29.0%
15	I thought of the ideas only in ENGLISH.	33.3%	25.1%	29.0%

*As respondents only had to answer Yes or No to this item, % agreement is based on those who answered 'yes'.

Table 8: Micro-planning (Agreement with questions 16–19)

Question		agree or strongly agree		
		Data (n=15 for Q16–18) (n=29 for Q19)	Diagram (n=16 for Q16–18) (n=31 for Q19)	Total (n=31 for Q16–18) (n=60 for Q19)
16	I was able to prioritise the ideas.	40.0%	37.6%	38.7%
17	I was able to put my ideas or content in good order.	46.7%	31.1%	38.8%
18	Some ideas had to be removed while I was putting them in good order.	40.0%	37.6%	38.7%
19	I felt it was easy to put ideas in good order.	17.2%	58.1%	38.4%

Table 9: Translating (Agreement with questions 20–26)

Question		agree or strongly agree		
		Data (n=29)	Diagram (n=31)	Total (n=60)
20	I felt it was easy to express ideas using the appropriate words.	17.2%	45.1%	31.7%
21	I felt it was easy to express ideas using the correct sentences.	24.1%	25.8%	25.0%
22	I thought of MOST of my ideas for the summary WHILE I was actually writing it.	41.3%	64.6%	53.3%
23	I was able to express my ideas by using appropriate words.	13.8%	61.3%	38.4%
24	I was able to express my ideas using CORRECT sentence structures.	20.6%	45.2%	33.3%
25	I was able to develop any paragraph by putting sentences in logical order in the paragraph.	31.0%	64.5%	48.3%
26	I was able to CONNECT my ideas smoothly in the whole response.	13.8%	41.9%	28.4%

Translating

It is at this stage that decisions made at macro-planning move from the abstract to the concrete where the writer encodes their ideas into the written form. It is at this stage where the L2 writer may face particular problems dependent on their language resources. The questions in Table 9 therefore relate to the ease or otherwise of this process of conversion of abstract ideas (mostly thought in L1) to the linguistic form of L2.

An initial glance would suggest that most of this sample did not find the translating stage easy. They also did not think that they were able to express their ideas appropriately and accurately. For questions 20 and 21 the levels of disagreement were quite high – 53.4% and 45% respectively – which suggests that this cohort was not particularly proficient or confident in their lexical and grammatical knowledge of English.

The figures were only marginally higher for questions 23 and 24 which focused on the ability as opposed to the ease with which they expressed their ideas lexically and grammatically. For both questions it was those who responded to the diagram task that expressed markedly higher levels of agreement suggesting that these candidates found the data task much more challenging lexically. This was perhaps partly due to the lack of lexical support when compared with the brick task (e.g. digger, clay, mould, etc. are on the question paper). Regarding responses to questions 25 and 26 just under 50% (and 64.5% of the diagram respondents) thought they were able to connect ideas within each paragraph but far fewer felt that they were able to organise the information as a whole (28.4%). Only 13.8% of the data respondents expressed agreement which suggests that building a coherent response to a graph showing quite a number of variables may be more challenging than a process task where the structure is almost self-evident.

Monitoring and revising

When a writer reviews at the sentence, paragraph or whole text level this involves the process of monitoring. If writing at any of these levels is found unsatisfactory, the writer is likely to revise, which could involve correcting a typographical error at one extreme to a wholesale re-draft of the whole text at the other. So although the socio-cognitive model presents these as two separate stages, these processes are so inextricably linked that for the purposes of analysis questions 27 to 38 were used to elicit information on both types of revision (see Table 10).

Monitoring is a very demanding activity so it is likely that the lower-level checking of mechanical accuracy of spelling, punctuation and syntax (questions 32–35) will exceed the higher-level checking of how the text fits in with the goals established in macro-planning and the text produced so far (questions 28–31). The figures below, however, do not seem to bear this out with both types of monitoring exhibiting fairly similar levels of agreement.

Questions 27 and 36–38 show much lower levels of agreement. These focus more on revision after the text as a whole has been written. Only 28% tried to take into account the word count (question 27) constraints or wrote a redraft (question 36). Slightly more reviewed any statements or thoughts that they had removed (33% in question 37) and just over a quarter of these candidates thought it easy to review and revise the whole response. It seems that as the rubrics recommend just 20 minutes for the completion of Task 1 it is time constraints that are probably the main factor in these low levels of agreement.

An interesting finding from the questionnaire data is the degree of difference in agreement between those responding to the data and the diagram task, although as the numbers involved are quite low any conclusions must be treated as very tentative. With 29 and 31 respondents respectively there is, however, a statistical procedure that could be used to see if there was any significant difference between the two tasks.

From sampling the distribution of differences between means a t-test for independent samples with equal variance revealed no significant difference in the distribution of differences between means between the two groups ($t=1.792$, $df=72$, $p=0.005$). Thus there is no evidence to suggest that the means between the two groups are different across the two task types, indicating that there is little difference in the perception of candidates between these two task types.

Table 10: Monitoring and revising (Agreement with questions 27–38)

Question		agree or strongly agree		
		Data (n=29)	Diagram (n=31)	Total (n=60)
27	I tried NOT to write more than the required number of words in the instructions.	31.0%	25.8%	28.3%
28	I reviewed the correctness of the contents and their order WHILE writing this response.	44.8%	45.1%	45.0%
29	I reviewed the correctness of the contents and their order AFTER finishing this response.	44.8%	48.4%	46.6%
30	I reviewed the appropriateness of the contents and their order WHILE writing this response.	41.4%	45.2%	43.3%
31	I reviewed the appropriateness of the contents and their order AFTER finishing this response.	48.3%	42.0%	45.0%
32	I reviewed the correctness of sentences WHILE writing this response.	51.8%	51.7%	51.6%
33	I reviewed the correctness of sentences AFTER finishing this response.	44.8%	41.9%	43.4%
34	I reviewed the appropriateness of words WHILE writing this response.	51.7%	54.8%	53.4%
35	I reviewed the appropriateness of words AFTER finishing this response.	44.8%	41.9%	43.4%
36	I was able to write a draft response in this test, then wrote the response again neatly within the given time.	37.9%	19.4%	28.4%
37	After finishing the summary I also thought for a while of those statements or thoughts I removed.	37.9%	29.0%	33.4%
38	I felt it was easy to review or revise the whole response.	24.1%	29.0%	26.7%

Conclusion and recommendations

For the cognitive processes required to complete the Task 1 in IELTS Academic Writing to be deemed appropriate, they need to replicate those thought processes that test takers will need to utilise in the future target language use situation. This study demonstrates that there is evidence of a large variety of the cognitive processes being employed, although organising does not seem to be as activated as much as the other processes. This is perhaps because ultimately the completion of Task 1 requires a knowledge-telling strategy even with very proficient writers. Unskilled writers are likely to plan less with each sentence generating the content of the next piece of text in a linear non-reflective manner. Skilled writers on the other hand may find that re-shaping the content from a visual input is not particularly demanding. They may adopt problem-solving strategies involved in knowledge transformation such as organising, but knowledge telling may be successful with very straightforward Task 1s.

In order to follow up this study and to furnish further evidence of cognitive validity to support the use of IELTS Academic Writing Task 1 the following research projects could be initiated:

- Further verbal protocol analysis where each informant would verbalise their thoughts on both data and diagram input tasks. Comparisons were limited in my study as the task variable was confounded by the participant variable.
- Keystroke logging of responses during VPA as subjects type their responses. This kind of research will become increasingly relevant as the IELTS partners plan to offer computer-based variations on the traditional pen and paper administrations they currently offer. Keystroke logging provides a more accurate record of when and where writers pause and together with concurrent protocols potentially offers richer data.
- Analysis of linguistic features of scripts from the VPA participants to gain further insight into levels of processing in terms of rhetorical and content parameters.

IELTS has always been a research-led enterprise and so these and other studies are likely to come to fruition in one form or another. As a high-stakes test it is important that IELTS continues to demonstrate validity. It is hoped that this small scale study using a relatively recent theoretical framework contributes in some way to the validity argument supporting the use of IELTS as a means of assessing the writing ability of those wishing to study or work in the medium of English.

References

Bachman, L (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

Bridges, G (2008) *Demonstrating further evidence of cognitive and context validity for Task 1 of the IELTS Academic Writing Paper using a socio-cognitive validity framework*, unpublished MA dissertation, Anglia Ruskin University.

Chapelle, C (1998) Construct definition and validity inquiry in SLA research, in Bachman, L and Cohen, A (Eds) *Second Language acquisition and language testing interfaces*, Cambridge: Cambridge University Press, 32–70.

Ellis, R (1994) *The Study of Second Language Acquisition*, Oxford: Oxford University Press.

Eysenck, M and Keane, M (2005) *Cognitive Psychology* (5th edition), Hove: Psychology Press.

Field, J (2004) *Psycholinguistics: the Key Concepts*, London: Routledge.

Green, A (1998) *Verbal protocol analysis in language testing research*, Cambridge: UCLES/Cambridge University Press.

Hyland, K (2002) *Teaching and Researching Writing*, London: Longman.

IELTS Scores Explained DVD (2006), Cambridge: Cambridge ESOL Publications.

Mayor, B, Hewings, A, North, S, Swann, J and Coffin, C (2006) A linguistic analysis of Chinese and Greek L1 scripts for IELTS Academic Writing Task 2, in Taylor, L and Falvey, P (Eds) *IELTS Collected Papers: Research in speaking and writing assessment*, Cambridge: Cambridge ESOL/Cambridge University Press, 250–315.

Mickan, P, Slater, S and Gibson, C (2000) Study of Response Validity of the IELTS Writing Subtest, in Tulloh, R (Ed.) *IELTS Research Reports Volume 3*, Canberra: IELTS Australia, 29–48.

Moore, T and Morton, J (2006) Authenticity in the IELTS Academic Writing test: a comparative study of Task 2 items and university assignments, in Taylor, L and Falvey, P (Eds) *IELTS Collected Papers: Research in speaking and writing assessment*, Cambridge: Cambridge ESOL/Cambridge University Press, 197–249.

Saville, N (2003) The process of test development and revision within UCLES EFL, in Weir, C and Milanovic, M (Eds) *Continuity and innovation: revising the Cambridge Proficiency in English Examination 1913–2002*, Cambridge: UCLES/Cambridge University Press, 57–120.

Scardamalia, M and Bereiter, C (1987) Knowledge telling and knowledge transforming in written composition, in Rosenberg, S (Ed.) *Advances in Applied Psycholinguistics, Volume 2: Reading, writing and language learning*, Cambridge: Cambridge University Press, 142–175.

Shaw, S and Khalifa, H (2007) *Deconstructing the Main Suite tests to understand them better*, Cambridge ESOL presentation to internal staff.

Shaw, S and Weir, C (2007) *Examining Writing: Research and practice in assessing second language writing*, Cambridge: Cambridge ESOL/Cambridge University Press.

Taylor, L and Falvey, P (2007) *IELTS Collected Papers: Research in speaking and writing assessment* Cambridge: Cambridge ESOL/Cambridge University Press.

Weir, C (2005) *Language Testing and Validation: an evidence-based approach*, Basingstoke: Palgrave Macmillan.

Weir, C, O'Sullivan, B, Jin Yan and Bax, S (2007) Does the computer make a difference? Reaction of candidates to a computer-based versus a traditional hand-written form of the IELTS Writing component: effects and impact, in Taylor, L (Ed.) *IELTS Research Report Volume 7*, IELTS Australia and British Council, 311–347.

Qualification and certainty in L2 writing: A learner corpus study

SIAN MORGAN CAMBRIDGE ESOL ORAL EXAMINER, UNIVERSITY OF MODENA AND REGGIO EMILIA, ITALY

Summary

This paper is based on a Master's thesis in TESOL submitted to Sheffield Hallam University (UK) in 2006. The research was funded by Cambridge ESOL. The MA was supervised by Dr Mary Williams.

The ability to express qualification and certainty is considered to be an important interpersonal skill which enables writers to avoid absolute statements and express caution in anticipation of criticism. Acknowledging the existence of possible alternative voices (Hyland 2005) plays a central role in building reader–writer relationships. It is important therefore that second language learners acquire flexible control of this skill in order for their writing to be successful. This paper describes a classroom research project carried out with second year language students at the University of Modena and Reggio Emilia (for more information see Morgan 2008). A small corpus of argumentative writing was compiled and examined to explore how this student population expressed qualification and certainty in their writing. The findings mirror those from previous studies of L2 writers: the students in this study rely on a small pool of modal verbs, overuse informal devices typical of spoken discourse, and tend to overstate their commitment to propositions. Implications for second language (L2) writing pedagogy and testing are discussed and some suggestions for consciousness-raising activities and form-focused practice are given.

Introduction

In recent years Corpus Linguistics (CL) has allowed us to examine authentic native English and observe linguistic and lexical patterns which occur typically in different writing contexts and discourse communities. This '*new perspective on the familiar*' (Hunston 2002:3) can also have useful applications in language teaching and pedagogy. One developing field of enquiry in corpus linguistics is the analysis of Computer Learner Corpora (CLC), which allows us to assemble authentic learner output and compare it to authentic, native-speaker (NS) data from a similar field or domain.

Such comparison can highlight what kind of features occur in L2 writing, and which of these occur most frequently. It can also give us information on *misuse*: what errors occur typically at which level. Equally interesting are the insights it gives us about the phenomenon of *under-use*, which does not lead to errors, but to under-representation of words or structures (Van Els, Bongaerts, Extra, van Os & Janssen-van Dieten 1984:63). By observing items which are avoided or distributed differently to comparable NS language, we are able to get a picture of aspects of language which present difficulties for specific groups of learners at different points on the interlanguage continuum. This information can yield insights about range, complexity and typical performance at different proficiency levels. In fact, CLC is currently being used in the English Profile project to describe in more detail linguistic and lexical features of learner output (McCarthy 2009). With C1 and C2 levels, where advanced language performance may reveal clusters of different features (Jarvis, Grant, Bikowski & Ferris 2003:399), corpus analysis may help us understand how these are distributed over student populations. Regardless of level however, if we are able to identify typical errors or avoidance strategies which still need to be addressed, we can then try to feed work on these areas into our teaching.

Focus of the study

This study was prompted by a previous investigation by Hyland and Milton (1997) into the way Hong Kong students express qualification and certainty in their writing. The authors believe that flexible use of linguistic devices to mitigate and boost statements is crucial to academic discourse for the following reasons:

Mitigators or 'hedges' allow writers to:

- avoid absolute statements
- acknowledge the presence of alternative voices
- express caution in anticipation of criticism.

Amplifiers or 'boosters' allow writers to:

- demonstrate confidence and commitment in a proposition
- mark their involvement and solidarity with the reader.

My own experience of working with Italian students suggests that they have firm control of amplifiers but are less likely to mitigate their statements. For example, several years ago one student, Chiara[1], wrote a well-structured and supported, generally accurate essay on the subject of teenage pregnancies in Britain, and was disappointed at receiving a slightly lower mark than she had expected. This was because she had failed to navigate the 'area between Yes and No' (Halliday 1985:335), and used only categorical statements with inappropriate strength of claim, resulting in what Milton (1999:230) has called 'over zealous emphasis'. If Chiara had qualified her statements more, in order to 'recognise alternative voices' (Hyland 2005:52) her essay would have been more persuasive. According to Hyland (2005:24):

> '… meaning is not synonymous with 'content' but dependent on **all** the components of a text. …both propositional and metadiscoursal elements occur together … each element expressing its own 'content': one concerned with the world, and the other with the text **and its reception**.' (bold added)

Equally importantly, as well as its central function in establishing the tone and style of academic writing, the ability to express qualification and certainty is considered to be an important politeness strategy in speech and writing. Salager-Meyer (1995) considers hedges and boosters to be '*a significant communicative resource for student writers at any proficiency level'*. Hyland & Milton (1997:186) also comment on this important area of pragmatic competence, and argue that these devices influence the reader's assessment of '*both referential and* ***affective*** *aspects of texts*' (bold added). In spoken discourse too, increasing attention has been paid to the pragmatic importance of hedging strategies. Carter (2005:68) suggests that they have an important interpersonal function in keeping lines of communication open; Hyland (2005) refers to this elsewhere as '*opening up a discursive space*' in written discourse.

All of this seems to suggest that flexible use of modal devices is important both as an interpersonal feature and as a communication strategy in L2 production in general. It is because of their all-pervasive nature in many types of discourse, as well as their significance in academic writing, that I decided to carry out a preliminary study using learner corpora to investigate the frequency and occurrence of these devices in my own local teaching context.

The research question was the following: how do undergraduate students express qualification and certainty in their argumentative writing, and what type of devices do they use most frequently?

Student profile and methods

Although the learner corpus used is very small, Granger (1998a) suggests that small corpora compiled by teachers of their own students' work can yield useful insights into a group profile of learner language. Clearly, for any corpus to be useful it is essential to have clear design criteria; in the case of learner language it is particularly important to control for the many different types of learner language and situations, taking into account variables such as the following:

Table 1: Variables to control for in learner corpora design

Language	Learner
medium	age
genre	sex
topic	L1
task	level
task setting	learning context

(Adapted from Granger 1998b:9)

The students in this project formed a relatively homogenous group in terms of age, level and language learning background. The 50 students involved were in their second year of a degree in European languages and culture at the University of Modena and Reggio Emilia. This was a predominantly female student population (42 female and 8 male) whose language level ranged from high B2 to low C1, as measured by their results in the first year exam. The study was conducted with this group of high-intermediate students as it was hoped that their firm

[1] A pseudonym

control of grammatical and lexical resources would free them up to reflect upon how modal or epistemic devices could be used to achieve different rhetorical purposes. The aim was to observe how they hedged or boosted their statements; therefore the focus here was on appropriateness, rather than accuracy.

The data used are two small corpora based on student writing produced at the end of the first and second semesters. CORPUS 1 was compiled of two short argumentative writing tasks submitted in the first semester. The handwritten scripts were later keyed into the computer verbatim by the students themselves. I then corrected typographical errors only and analysed the texts using Wordsmith Tools text retrieval software to examine the type and frequency of hedges and boosters occurring in the scripts. A further manual analysis was conducted to disambiguate any items. CORPUS 2 was compiled from two further assignments submitted at the end of the second semester and a similar analysis was carried out.

Findings and analysis

In order for a learner corpus to be meaningful it needs to be compared to some kind of norm. For this, I used Hyland & Milton's (1997:196) taxonomy of the most frequently appearing epistemic items in academic discourse, and observed which of these items occurred in these two learner corpora.

Informal items

The students in this population also used a considerable number of informal items which were not cited in Hyland & Milton's (1997) taxonomy (see Table 2).

Table 2: Top 10 epistemic devices which occurred in this study

CORPUS 1	No.	%	CORPUS 2	No.	%
can	112	4.8%	all	63	2.1%
all	73	3.2%	can	38	1.3%
everyone	27	1.2%	every	21	0.7%
every	26	1.2%	especially	20	0.7%
really	20	0.9%	according to	19	0.6%
in my opinion	19	0.8%	in my opinion	16	0.5%
especially	19	0.8%	sort of	13	0.4%
must	17	0.7%	really	12	0.4%
extremely	16	0.7%	show	11	0.4 %
completely	12	0.5%	completely	10	0.3%

This mirrors previous findings (Hinkel 2005, Hyland & Milton 1997, Milton 1999) which suggest that L2 writers rely more on items from spoken language and conversational discourse. For example, in this particular study, there is an overuse of informal items such as *really* which functions as an intensifier:

… not obligatory, they are **really** important.
… to find people who **really** like travelling and …
… met outside of school will **really** help you in your
… it doesn't **really** concern only …
… on these facts to be **really** effective; teachers …
… so work experience can **really** help you to grow…
… that's why you're **really** interested on it.
… world of sport has **really** changed today …
… the meeting are **really** serious and …
… have turned out to be **really** appreciated …

Expert NS and non-native-speaker (NNS) writers, in a similar argumentative task, might have achieved this emphasis more formally, for example, by replacing *really important* with *crucial, really help you* with *be of considerable help,* and *really appreciated* with *very much appreciated*.

Predominance of central modals

The same central modal verbs *will*, *should*, *would*, *could*, and the epistemic verb *think*, appeared in the top 10 tokens of both Corpus 1 and Corpus 2 (see Table 3).

Table 3: Occurrence of central modals in Corpus 1 and 2 (raw figures)

	CORPUS 1	CORPUS 2
No. words	23,470	29,560
No. texts	99	75
could	60	30
couldn't	1	0
may	7	9
might	5	4
should	39	20
shouldn't	0	0
would	81	29
wouldn't	0	0
will	123	31
won't	0	2
Total	**316**	**125**

Again this mirrors previous findings (Hinkel 2005, Hyland & Milton 1997) that both NS and NNS use the same pool of items in their writing, albeit with different frequency patterns. This may be partly developmental or interlingual; it may also be a result of teaching, or the large amounts of attention devoted to these items in textbooks (Hyland & Milton 1997:189). It does, however, seem to suggest that modal verbs are more automatically retrievable or easier to manipulate for NNS writers than lexical modal devices, modal nouns or adverbs.

Epistemic verbs

After central modals, the next most frequent items were epistemic verbs such as *think*, *know* and *believe*, together with usuality markers such as *always* and *usually*. Hinkel's (2005) finding that *think* rather than *believe* is preferred by NNS writers is replicated here with *think* appearing in third and second position in Corpus 1 and 2 (Table 4 overleaf).

Several studies have confirmed this overuse of *I think* as a popular sentence builder in L2 writing, occurring three to five times more frequently in NNS writing compared to NS writing (Granger 1998a).

Table 4: Top 10 modal devices occurring in Corpus 1 and 2 (percentages)

CORPUS 1	No.	%	CORPUS 2	No.	%
will	123	5%	in fact	41	1.4%
would	81	3.5%	think	34	1.2%
think	61	2.7%	will	31	1.2%
could	60	2.6%	could	30	1.1%
should	39	1.7%	would	29	0.9%
always	26	1.1%	always	29	0.9%
in fact	24	1%	quite	28	0.9%
know	19	0.8%	should	20	0.7%
usually	18	0.8%	clear	16	0.5%
possible	15	0.6%	believe	15	0.5%

Predominance of boosters

It is also interesting to note that boosters (which have an amplifying function) rather than hedges (mitigating function) predominate in the list of 10 most frequently occurring devices in this corpus. This may be a result of a mother tongue (L1) fingerprint on L2, although this hypothesis would need to be researched further for an Italian L1 context. Past learning experience or instruction where students are encouraged to express their views assertively may also be a contributory factor.

Sentence position

Previous studies of complexity in L2 writing have found that, possibly because of the multiple demands of the composing process, learners frequently default to safe usages such as *thing* instead of *topic issue/question*. In this corpus, too, the same phenomenon occurs when expressing opinions. For instance, many students in this corpus relied on personal subjectivity markers such as *in my opinion*, what Hasselgren (1994) might describe as a 'lexical teddy bear'. This is illustrated in the examples below:

> … instead of having a walk with a friend. **In my opinion**, it would be better spend …
>
> … "real" encounter takes place. **In my opinion**, to deal with this issue …
>
> … action proposing these two projects. **In my opinion**, proposal number one is …
>
> … the Car Park and the city centre, yet **in my opinion** this may be revealed as …
>
> … threatened or highly endangered. **In my opinion**, we have led our planet …
>
> The first proposal is, **in my opinion**, a great solution for …
>
> … and stressful sport activity. **In my opinion** the secret for staying fit …
>
> … the health side to doing sports. **In my opinion** practicing sports, and …
>
> … too much traffic and much noise. **In my opinion** a good solution for …

It is also interesting to observe where these hedges are used, and to speculate whether positioning can strengthen writer commitment. In this corpus, *in my opinion* occurs often at sentence-initial position, in contrast to NS writing where it occurs frequently in a subordinate or clause-initial position. It is arguable whether *in my opinion* at sentence-initial position has a mitigating or amplifying effect on writer commitment, and if this effect might change if it were embedded or inserted at clause-initial position.

It may be that NNS writers prefer to use fixed phrases in sentence-initial position because they are often presented in school textbooks in this way, and this makes them implicitly available for uptake by students. This might be something we want to draw students' attention to when using published materials.

Compound hedges

Despite the predominance of boosters in this learner corpus, there were also some clear attempts to qualify assertions. For example, some students tried to combine devices in a 'compound hedge' (Salager-Meyer 1995:155), not always with harmonic results. Nevertheless, it is interesting to note that such clusters, typical of expert or NS writers, also occurred in this corpus (see examples below). This seems to indicate an increasing awareness of the reader–writer relationship in this high-intermediate student population.

> **If it is possible for me to make a suggestion**, my advice **would be** to try to reduce the number of cars circulating
>
> …. **or rather I would say that I feel the need to express my opinion** concerning …
>
> **Personally, I think that** imposing a daily "congestion charge" **could be a good idea**….
>
> This restriction **seems to me not quite** right …

Some researchers (e.g. Hyland & Milton 1997) have found that students who modify their statements with more tentative expressions tend to have a higher level of general language proficiency. Others, instead, suggest that although greater linguistic competence is an important pre-requisite, it does not automatically imply the parallel development of pragmatic competence (Bardovi-Harlig & Dörnyei 1998:234).

Possible reasons for lack of control of modal devices

Even in this small study of a relatively homogenous student population there was some variation both in the degree of formality and the degree in the use of tentativeness. This may be linked to one or more of the following factors:

- language level (even within this relatively homogeneous student population)
- writing competence (as opposed to language competence)
- incomplete register control
- individual differences in communicative style
- cultural differences in rhetorical style.

Discussion and implications for teaching and testing

The findings of this initial experiment indicate that many high-intermediate students in this study used modality markers to express qualification and certainty. Like other populations previously studied (Hinkel 2005, Hyland & Milton 1997), they tend to overstate rather then hedge their assertions, possibly in a bid to 'sell' their ideas, and often default to informal items (e.g. *really*), creating a degree of writer visibility which may not be appropriate in all types of writing. Also, the narrow range of modal auxiliaries which learners tend to rely on at this stage may not be adequate as they progress to more complex, pragmatically sensitive writing events in future contexts, both academic and professional. Therefore it is important to make learners aware that there is a wider spectrum of linguistic choices available for these purposes and to provide opportunities for them to encounter such alternatives in context.

A further consideration is the improvement of stylistic proficiency, which is an important objective as students progress along the writing continuum. The increasing internationalisation of higher education means that, in order to gain access to English-medium university courses, students need to obtain advanced English language qualifications such as International English Language Testing System (IELTS), CAE (Cambridge English: Advanced[2]) or CPE (Certificate of Proficiency in English). Testing criteria for these exams, based on the Common European Framework of Reference (CEFR) descriptors, include lexical resources and interactive communication. To meet the required level for C1 and C2, students need to use a wide range of lexis accurately and appropriately to perform interpersonal functions and meet the testing criteria. Therefore a strong learner training component in exam preparation classes could provide learners with strategies to extend their range of lexis and discover alternatives to certain default usages or 'islands of reliability' (Dechert 1984:227).

For second language learners, increasing their stock of lexis is a particular challenge (Schmitt 2008:329). Research on advanced students' vocabulary (Ringbom 1998:43) has shown that learners at this level consistently use the 100 most frequent words more often than NS writers. Rundell & Granger (2007) report corpus findings demonstrating that learners writing academic texts use the discourse marker *besides* about 15 times more frequently than native speakers writing in the same mode. Such findings highlight how expanding lexical resources is a key priority for learners, and how vocabulary acquisition should concern not only content words, but also a range of lexis to perform interpersonal functions such as agreeing, disagreeing or expressing opinion. For example, the findings of this particular study suggest that these students need to develop their repertoire of alternatives to central modals. Sinclair's (1991) idiom (rather than open choice) principle holds that meaning is attached to the whole phrase rather than the individual parts of it, so teachers may want to draw students' attention to prefabricated modal chunks (lexical phrases) as they are encountered, as well as individual tokens (modal verbs).

As well as providing opportunities for intentional learning of vocabulary, we need to provide opportunities for incidental learning of vocabulary (Schmitt 2008:353). Students may benefit from exposure to appropriate text models through extensive reading of a variety of text types. In this way they can explore contextualised examples of these devices, notice how they occur typically in discourse, and reflect on their function in each context. For example, the predominance of hedges in the abstract and discussion section of an academic article are polypragmatic in that they express a degree of uncertainty and therefore humility towards the academic community. Apprentice texts written by advanced-level students (Flowerdew 2000) can also be an excellent source of reading texts for students of slightly lower levels. Attention can be drawn to hedging devices, which are often lexically invisible to learners (Lowe 1996:30), and the possible purpose of these can then be discussed. For example, they may be used to express caution in anticipation of criticism, to show politeness and modesty towards the academic community and wider readership, or to open up a dialogical space, among others.

The following are some suggestions for form-focused instruction and consciousness-raising (CR) activities:

- remove hedges from texts and ask students to discuss the resulting effect on the reader
- ask students to explore the function of multi-word items which naturally occur in the target discourse such as *it would seem that*, *to my knowledge*, *to some extent* or the more informal *on the whole* in their reading (and notice that they are sometimes embedded in the clause and not in sentence-initial position)
- ask students to distinguish statements in a text which report facts and those which are unproven
- students rewrite an academic essay (which uses hedges and boosters) into popular journalistic style (which doesn't) or vice versa (Hyland 2005)
- design persuasive tasks of various kinds on sensitive topics, anticipating the potentially critical views of the reader (Hyland 2005)
- students could reformulate texts to accommodate different audiences, and compare the before and after effect on the audience.

Conclusion

This has been a preliminary investigation into an area of learner language which is receiving increasing attention from discourse analysts. The study should be regarded as a point of departure rather than arrival, and the findings are intended to be representative of a specific student population only. Clearly, it would benefit from further quantitative and qualitative analysis and replication in other student populations. Nevertheless, it has thrown up interesting insights about how the students in this setting navigate the 'area of meaning between Yes and No', which I have since used to inform my teaching. What it suggests is that we may need to adopt a more systematic approach to raising students' awareness of these interpersonal features in building reader–writer relationships and fostering effective communication in general. In this way, unlike Chiara in her essay on teenage pregnancies, they can learn to acknowledge the presence of 'alternative voices'.

[2] Previously known as Certificate in Advanced English

References and further reading

Bardovi-Harlig, K and Dörnyei, Z (1998) Do language learners recognize pragmatic violations? Pragmatic versus grammatical awareness in instructed L2 learning, *TESOL Quarterly* 32 (2), 233–262.

Carter, R (2005) *What is a frequent word?,* paper presented at the international IATEFL conference, Cardiff, 5–9 April, 2005.

Dechert, H (1984) Second language production: Six hypotheses, in Dechert, H, Mohle, D and Raupach, M (Eds) *Second Language Productions*, Tubingen: Gunter Narr Verlag, 211–223.

Flowerdew, L (2000) Using a genre-based framework to teach organizational structure in academic writing, *ELT Journal* 54 (4), 369–378.

Granger, S (1998a) Prefabricated patterns in advanced ELT writing: collocations and formulae, in Cowie, A P (Ed.) *Phraseology: theory, analysis, and applications*, Oxford: Clarendon Press, 145–160.

Granger, S (1998b) The computer learner corpus: a versatile new source of data for SLA research, in Granger, S (Ed.) *Learner English on Computer*, New York: Pearson Education, 3–18.

Granger, S (2002) A Bird's eye view of Learner Corpus Research, in Granger, S, Hung, J and Petch-Tyson, S (Eds) *Computer Learner Corpora, Second Language Acquisition and Foreign Language teaching*, Amsterdam/Philadelphia: John Benjamins Publishing Company, 3–33.

Halliday, M (1985) *An Introduction to Functional Grammar*, London: Arnold.

Hasselgren, A (1994) Lexical teddy bears and advanced learners: a study into the way Norwegian students cope with English vocabulary, *International Journal of Applied Linguistics* 4, 237–58.

Hinkel, E (2005) Hedging, inflating and persuading, *Applied language learning* 15 (1–2), 29–53.

Hunston, S (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.

Hyland, K (2000) Hedges, Boosters and Lexical Invisibility: Noticing Modifiers in Academic Texts, *Language Awareness* 9 (4), 179–197.

Hyland, K (2005) *Metadiscourse*, London/New York: Continuum.

Hyland, K and Milton, J (1997) Qualification and Certainty in L1 and L2 Students' Writing, *Journal of Second Language Writing* 6 (2), 183–205.

Jarvis, S, Grant, L, Bikowski, D and Ferris, D (2003) Exploring multiple profiles of highly rated learner compositions, *Journal of Second Language Writing* 12, 377–403.

Lowe, G (1996) Intensifiers and Hedges in Questionnaire items and the Lexical Invisibility Hypothesis, *Applied linguistics*, 17 (1), 1–37.

McCarthy, M (2009) *English Profile*. TESOL Talk from Nottingham, retrieved from http://portal.lsri.nottingham.ac.uk/SiteDirectory/TTfN/default.asp

Milton, J (1999) Lexical thickets and electronic gateways, in Candlin, C N and Hyland, K (Eds) *Writing: texts, processes and practices*, London: Longman, 221–244.

Morgan, B S (2008) The space between Yes and No: how Italian students qualify and boost their statements, in Palawek, M (Ed.) *Investigating English Language Learning and Teaching*, Poznan-Kalisz: Adam Mickiewicz University, 267–278.

Ringbom, H (1998) Vocabulary frequencies in advanced learner English: A cross-linguistic approach, in Granger, S (Ed.) *Learner English on Computer*, London & New York: Addison Wesley Longman, 41–52.

Rundell, M and Granger, S (2007) From Corpus to confidence, *retrieved from:* http://www.macmillandictionaries.com/MED-Magazine/August2007/46-Feature_CorporatoC.htm

Salager-Meyer, F (1995) I Think That Perhaps You Should: A Study of Hedges in Written Scientific Discourse, *Journal of TESOL France* 2, 127–143.

Schmitt, N (2008) Review article: Instructed second language vocabulary learning, *Language Teaching Research* 12, 329–363.

Sinclair, J M (1991) *Corpus Concordance Collocation*, Oxford: Oxford University Press.

Van Els, T, Bongaerts, T, Extra, G, van Os, C, and Janssen-van Dieten, A M (1984) *Applied Linguistics and the Learning and Teaching Languages*, Edward Arnold: London.

Prompt and rater effects in second language writing performance assessment

GAD S LIM RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

This short summary is based on a doctoral thesis submitted to the University of Michigan, Ann Arbor (US) in 2009. The PhD was supervised by Professor Diane Larsen-Freeman.

Performance assessments have become the norm for evaluating language learners' writing abilities in international examinations of English proficiency. Two aspects of these assessments are usually systematically varied: test takers respond to different prompts, and their responses are read by different raters. This raises the possibility of undue prompt and rater effects on test takers' scores, which can affect the validity, reliability and fairness of these tests.

This study uses data from the Michigan English Language Assessment Battery (MELAB), including all official ratings given over a period of over four years (n=29,831), to examine these issues related to scoring validity. It uses the multi-facet extension of Rasch methodology to model this data, producing measures on a common, interval scale. First, the study investigates the comparability of prompts that differ on topic domain, rhetorical task, prompt length, task constraint, expected grammatical person of response, and number of tasks. It also considers whether prompts are differentially difficult for test takers of different genders, language backgrounds, and proficiency levels. Second, the study investigates the quality of raters' ratings, whether these are affected by time and by raters' experience and language background. It also considers whether raters alter their rating behaviour depending on their perceptions of prompt difficulty and of test takers' prompt selection behaviour.

The results show that test takers' scores reflect actual ability in the construct being measured as operationalised in the rating scale, and are generally not affected by a range of prompt dimensions, rater variables, test taker characteristics, or interactions thereof. It can be concluded that scores on this test and others like it have score validity and, assuming that other inferences in the validity argument are similarly warranted, can be used as a basis for making appropriate decisions. Further studies to develop a framework of task difficulty and a model of rater development are proposed.

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

LUCY CHAMBERS RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

This short summary is based on a Master's thesis submitted to the Faculty of Arts, Law and Social Sciences, Anglia Ruskin University in 2007. The research was funded by Cambridge ESOL. The MA was supervised by Dr Sebastian Rasinger.

This MA research focused on Cambridge ESOL's Preliminary English Test (PET).

In 2007 Cambridge ESOL was starting to launch computer-based versions of many of its paper-based tests. Thus it was important that the issues of comparability between administration modes were explored. This study focuses on the skill of writing and builds on research from overall score and writing sub-element score comparability studies. Unlike the majority of current research, which focuses on score comparability, this study focuses on the comparability of text and linguistic features. Features studied include lexical range and sophistication, text length and organisation and surface features such as capitalisation and punctuation.

The study is set within an ESOL assessment environment and is in two parts. The first part is a qualitative analysis of a small sample of scripts that also acts as a pilot for part two. Tasks from Cambridge ESOL's Preliminary English Test (PET) are used and the resulting scripts from paper-based and computer-based administrations analysed.

In the second and main part of the study scripts produced from a live PET administration were studied. Two samples of texts were chosen; these samples were matched on candidates' proficiency and the country in which they sat the exam. A number of linguistic and text features were analysed. Texts were found to be comparable in text length, surface features and lexical error rates. However, there were differences in lexical variation and in the number of sentences and paragraphs produced. It is recommended that these results be considered a starting point from which to further explore text-level differences across writing modes, covering additional first languages, proficiency levels and writing genres. Results from this and future studies can help inform rater training and provide information for teachers and candidates. For more details on this study see Chambers (2008).

References

Chambers, L (2008) Computer-based and paper-based Writing assessment: a comparative text analysis, *Research Notes* 34, 9–15.

A study of the context and cognitive validity of a BEC Vantage Test of Writing

HUGH BATEMAN ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

This short summary is based on a Master's thesis submitted to Anglia Ruskin University in 2008. The research was funded by Cambridge ESOL. The MA was supervised Dr Sebastian Rasinger.

This study applied Weir's (2005) socio-cognitive framework to investigate context and cognitive validity of the Writing component of a test of English in a business context. Cognitive validity was investigated primarily through a small-scale, qualitative study which used verbal protocol analysis to establish whether one of the test tasks activated the same cognitive processes as similar tasks in the real-life workplace. Cognitive validity was found to be high. All three subjects displayed the same five stages of cognitive processing in completing the test task and the real-life task. However, there was no evidence in either task of a sixth stage identified in the above framework, in which writers organise ideas in a pre-linguistic form. It seems probable that the lack of an organisation phase is related to the brevity of the tasks rather than their English for Specific Purposes (ESP) nature. The fine-grained processing operations of all three subjects were very similar for both tasks in the translation and monitoring phases. Two of the three subjects displayed very similar micro-planning operations on both tasks; however, micro-planning of the third subject's test task was influenced by a desire not to exceed the word limit specified in the task. Consideration of the word limit also influenced one subject's macro-planning of the test task, and all three subjects engaged in considerably more macro-planning for the test task than their real-life task. However, there was no evidence that macro-planning was affected by completing the test task on paper rather than on computer. All three subjects engaged in similar revising activity on both tasks. There was no evidence that the limits on major revisions to wording or structure that apply when handwriting a test task resulted in different cognitive processing operations to a word-processed task. For a summary of the part of the study that investigated the test's context validity (specifically, the linguistic demands the test made of the candidates who took it), see Bateman (2009).

References

Bateman, H (2009) Some evidence supporting the alignment of an LSP Writing test to the CEFR, *Research Notes* 37, 29–34.

Weir, C J (2005) *Language Testing and Validation: An Evidence-Based Approach*, Basingstoke: Palgrave Macmillan.

Models of supervision – some considerations

JULIET WILSON CUSTOMER SERVICES GROUP, CAMBRIDGE ESOL

This short summary is based on the report submitted as part of requirements for an MA in TESOL at the University of London in 1996. The thesis was supervised by Dr John Norrish.

My report *Models of Supervision – some considerations* was concerned with aspects of teacher supervision. After summarising various historical approaches to teacher supervision and feedback, I outlined some of the factors which need to be taken into account when evaluating the potential of these different models, including the education/training debate, the issue of teacher evaluation and the resulting roles that a supervisor may be called upon to carry out. The report went on to consider two case studies – my experiences as a trainer on a pre-service certificate course at a Further Education College in London, and as a supervisor at a secondary school in Malta during the practicum of the Teacher Education and Training module of the MA. I explored the limitations and successes of these two experiences and showed how the work I did in Malta modified my view of the supervisory process and led me to draw some tentative conclusions about the advantages of a non-evaluative, co-operative approach to teacher training.

A framework for analysing and comparing CEFR-linked certification exams

MARYLIN KIES CENTRE EXAMINATION MANAGER, UNIVERSITY OF SIENA, ITALY

This short summary is based on a Master's dissertation submitted to the University of London Institute of Education in 2009. It was supervised by Dr Amos Paran.

Communication and transparency are fundamental ideals underlying the Council of Europe Common European Framework of Reference (CEFR). The CEFR has facilitated communication immensely, as teachers, students, publishers, policy makers and examination boards all now make reference to the CEFR levels. Transparency, however, presents a greater challenge, at least regarding language certification. Although test users may presume that exams pegged to the same CEFR level are 'in some way equivalent' or at 'exactly the same level' (COE 2009:4), this is not necessarily so, as the Council of Europe (COE) encourages diversity. Moreover, interpretation of the CEFR specifications varies considerably and no overseeing authority monitors claims of linkage. As students and aspiring employees normally choose the certification exams recognised, required or offered by institutions and employers, these latter must set their policies wisely.

This study suggested how institutional and professional test users may analyse and compare certification exams linked to the CEFR using three sets of criteria: Weir's (2005) socio-cognitive validity framework to evaluate overall test validity, the CEFR scales to evaluate the extent to which these are addressed in test tasks and the COE's (2009) *Manual* for relating language examinations to the CEFR to assess the validity of linkage to a CEFR level. To illustrate this procedure, two 4-skills B1 certification exams in English for speakers of other languages were compared: Cambridge ESOL's Preliminary English Test and Trinity College London's Integrated Skills in English 1. The resulting analysis revealed that even exams that are similar in terms of their characteristics, aims and recognition might not equally satisfy an institutional or professional test user's requirements.

References

Council of Europe, Language Policy Division (2009) *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) A Manual*, Strasbourg: Language Policy Division.

Weir, C J (2005) *Language Testing and Validation: An Evidence-Based Approach*, Basingstoke: Palgrave Macmillan.

IRT model fit from different perspectives

MUHAMMAD NAVEED KHALID RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

This short summary is based on a doctoral thesis submitted to the University of Twente (Netherlands) in 2010. The PhD was supervised by Professor Cees A W Glass.

The chapters in this thesis are self-contained; hence they can be read separately.

In Chapter 2, item bias or differential item functioning (DIF) is seen as a lack of fit to an IRT model. It is shown that inferences about the presence and importance of DIF can only be made if DIF is sufficiently modelled. This requires a process of so-called test purification where items with DIF are identified using statistical tests and DIF is modelled using group-specific item parameters. In the present study, DIF is identified using a Lagrange multiplier statistic. The first problem addressed is that the dependency of these statistics might cause problems in the presence of relatively large number DIF items. However, simulation studies show that the power and Type I error rate of a step wise procedure where DIF items are identified one at a time are good. The second problem pertains to the importance of DIF, i.e. the effect size, and related problem of defining a stopping rule for the searching procedure. Simulations show that the importance of DIF and the stopping rule can be based on the estimate of the difference between the means of the ability distributions of the studied groups of respondents. The searching procedure is stopped when the change in this effect size becomes negligible.

Chapter 3 presents the measures for evaluating the most important assumptions underlying unidimensional item response models such as subpopulation invariance, form of item response function, and local stochastic independence. These item fit statistics are studied in two frameworks. In a frequentist MML framework, LM tests for model fit based on residuals are studied. In the framework of LM model tests, the alternative hypothesis clarifies which assumptions are exactly targeted by the residuals. The alternative framework is the Bayesian one. The PPCs is a much used Bayesian model checking tool because it has an intuitive appeal, and is simple to apply. A number of simulation studies are presented that assess the Type I error rates and the power of the proposed item fit tests in both frameworks. Overall, the LM statistic performs better in terms of power and Type I error rates.

Chapter 4 presents fit statistics that are used for evaluating the degree fit between the chosen psychometric model and an examinee's item score pattern. Person fit statistic reflects the extent to which the examinee answered test questions according to the assumptions and description of the model. Frequentist tests as the LM test and tests with Snijders' correction (which take into account the estimation of ability parameter) are compared with PPCs. Simulation studies are carried out using number of fit statistics in a number of combinations in both frameworks.

In Chapter 5, a method based on structural equation modelling (or, more specifically, confirmatory factor analysis) for examining measurement equivalence is presented. Top-down and bottom-up approaches were evaluated for constructing nested models. A comprehensive comparative simulation study is carried out to explore the factors that have impact performance for detecting DIF items.

Conferences and publications

IACAT conference

LAURA COPE AND **TAMSIN WALKER** RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

The first conference of the International Association for Computerized Adaptive Testing (IACAT) was held from 7 to 9 June 2010 in Arnhem, the Netherlands. The conference was hosted by the Research Center for Examination and Certification (RCEC), a partnership between Cito and The University of Twente. Around 130 delegates from over 30 countries attended, representing a wide of range of interests: research, education, assessment, medical and commercial. Experience of CAT testing ranged from those who were attending due to an initial interest, to organisations which already employed CAT tests, to psychometricians who specialised in CAT.

Of the three workshops which were run on the first morning, Research and Validation attended 'Item Selection, Exposure Control, and Test Specifications in CAT', given by Bernard Veldkamp of RCEC. This described the use of linear programming – mathematically defining requirements as a function which is then solved for an optimal solution – to provide an optimal set of test items which conform to a set of test requirements. These requirements can be: quantitative, such as the item difficulty; categorical, such as the task type; or logical, for instance, sets of items which cannot be used in the same test. Requirements can be specified from item level through to multiple-test level. The workshop included practical exercises in the formal specification of requirements. Once the requirements are defined, software packages are able to provide solutions within a split second. For CAT tests, an optimal linear test, the 'shadow test' is assembled online after each candidate response. After taking into account those items already used, the next item is picked from this set of items rather than the whole item pool.

Brian Bontempo from Mountain Measurement, USA, delivered an interesting presentation entitled 'The theoretical issues that are important to operational adaptive testing', the emphasis of which was the requirement for more research on operational and practical solutions to CAT problems, rather than theoretical work which is not implementable. The presentation examined item pool usage and argued that despite all of the research that has been focused on designing exposure control mechanisms, item exposure is only important practically if item parameter drift occurs. The question 'how much exposure is too much?' was left unanswered, however, the need to continually monitor items after calibration to check for item parameter drift was emphasised. The need to maximise the usage of items both in terms of over and under-exposure was emphasised, since a lot of research has focused on the prevention of item over-exposure. The presentation made the point that lowering the difficulty of items can improve test performance (psychologically) – this was a popular theme across a number of presentations. Finally, the question 'what should you do if the computer crashes mid-test?' was contemplated, with possible suggestions being to restart the test, start from the point at which the test stopped and start the test again the next day.

A useful overview on 'How to make adaptive testing more efficient' was given in a keynote presentation by Wim van der Linden. He suggested individualising the start of a test using collateral information, such as using data from previous instruction, or asking the candidate for a self-rating assessment. The use of covariates such as response times can also speed up convergence of the ability estimate. Improving the efficiency of item writing by rule-based item generation (cloning) was suggested; the efficiency of item calibration can then be improved by pretesting item families, rather than all individual items. Approaches to the optimal assembly of item pools, such as item pool rotation (which helps the issues of both over and under-exposure), and the idea of creating a pool as a set of test forms, each of which meets test requirement constraints, were covered.

ALTE events

Participants from as far afield as Chile, Libya and Qatar, together with others from the Czech Republic, Denmark, Germany, Spain and the UK signed up for the ALTE summer testing courses which took place from 20 to 24 September, and from 27 September to 1 October. These courses were hosted by the Basque Government, ALTE's Basque member, at the Royal Academy of the Basque Language in Bilbao. The first course was an *Introductory Course in Language Testing* run by Professor Cyril Weir and Dr Lynda Taylor, and the second was an *Introduction to Testing Reading* run by Dr Hanan Khalifa and Dr Ivana Vidaković from Research and Validation.

Later in the year, ALTE's 39th meeting and conference will take place at the Charles University in Prague from 10 to 12 November. As at previous meetings, the first two days will include a number of Special Interest Group meetings, and workshops for ALTE members and affiliates, and the third day will be an open conference day for anyone with an interest in language testing. The theme of the conference is '*Fairness and Quality Management in Language Testing*' and the speakers at the conference will include Professor Antony Kunnan and Dr Piet van Avermaet, as well as Dr Neil Jones, Juliet Wilson and Mike Gutteridge from Cambridge ESOL. Juliet, Mike and Dittany Rose will also run workshops on the two days prior to the conference day.

Just prior to the Prague conference, ALTE is launching the first of its Tier 3 language testing courses with a 2-day course on *The Application of Structural Equation Modelling (SEM) in Language Testing Research* on 8 and 9 November. The course will be run by Dr Ardeshir Geranpayeh from Research and Validation. This is an advanced course in language testing (ALTE Tier 3) and is aimed at experienced and knowledgeable language testing professionals. The Tier 3 courses complement the Foundation Courses (Tier 1) and Introductory Courses (Tier 2) which are already well established. Following the conference, on 13 November, ALTE will continue its programme of Foundation Courses when Annie Broadhead will run a general introduction to language testing.

The call for papers for the ALTE 4th International Conference to be held in Kraków, Poland from 7 to 9 July 2011 is already open and will run until the end of January 2011. We encourage you to submit a paper for the conference, and reflecting ALTE's commitment to multilingualism, papers can be submitted in English, French, German, Italian, Polish and Spanish. The theme of the conference is '*The Impact of Language Frameworks on Assessment, Learning and Teaching viewed from the perspectives of policies, procedures and challenges*' and the plenary speakers are Professor Lyle Bachman, Professor Giuliana Grego Bolli, Dr Neil Jones, Dr Waldemar Martyniuk, Dr Michaela Perlmann-Balme and Professor Elana Shohamy.

For further information about these events and other ALTE activities, please visit the ALTE website – www.alte.org

Studies in Language Testing

The last 12 months have seen the publication of three more titles in the *Studies in Language Testing* series, published jointly by Cambridge ESOL and Cambridge University Press.

Volume 29, authored by Hanan Khalifa and Cyril J Weir, is entitled *Examining Reading: Research and practice in assessing second language reading*. This volume develops a theoretical framework for validating tests of second language reading ability. The framework is then applied through an examination of the tasks in Cambridge ESOL Reading tests from a number of different validity perspectives that reflect the socio-cognitive nature of any assessment event. The authors show how an understanding and analysis of the framework and its components can assist test developers to operationalise their tests more effectively, especially in relation to the key criteria that differentiate one proficiency level from another.

Key features of the book include: an up-to-date review of the relevant literature on assessing reading; an accessible and systematic description of the different proficiency levels in second language reading; and a comprehensive and coherent basis for validating tests of reading. This volume is a rich source of information on all aspects of examining reading ability. As such, it will be of considerable interest to examination boards wishing to validate their own reading tests in a systematic and coherent manner, as well as to academic researchers and graduate students in the field of language assessment more generally. This is a companion volume to the previously published *Examining Writing* (Shaw & Weir 2007).

Volume 31, co-edited by Lynda Taylor and Cyril J Weir, is entitled *Language Testing Matters: Investigating the wider social and educational impact of assessment – Proceedings of the ALTE Cambridge Conference, April 2008*. It explores the social and educational impact of language testing and assessment, at regional, national and international level, by bringing together a collection of 20 edited papers based on presentations given at the 3rd international conference of the Association of Language Testers in Europe (ALTE) held in Cambridge in April 2008.

The selected papers focus on three core strands addressed during the conference. Section One considers new perspectives on testing for specific purposes, including the key role played by language assessment in the aviation industry, in the legal system, and in migration and citizenship policy. Section Two contains insights on testing policy and practice in the context of language teaching and learning in different parts of the world, including Africa, Europe, North America and Asia. Section Three offers reflections on the impact of testing among differing stakeholder constituencies, such as the individual learner, educational authorities, and society in general.

Key features of the volume include: up-to-date information on the impact of language testing and assessment in a wide variety of social and educational contexts worldwide; accounts of recent research into the profiling of language proficiency levels and into cheating in tests; insights into new areas for testing and assessment, e.g. teacher certification, examinations in L2 school systems, testing of intercultural competence; discussion of the relationships among different test stakeholder constituencies.

With its broad coverage of key issues, combining theoretical insights and practical advice, this volume is a valuable reference work for academics, employers and policy-makers in Europe and beyond. It is also a useful resource for postgraduate students of language testing and for practitioners, i.e. teachers, teacher educators, curriculum developers, materials writers, and anyone seeking greater understanding of the social and educational impact of language assessment.

July 2010 saw the publication of another title in the *Studies in Language Testing* series, published jointly by Cambridge ESOL and Cambridge University Press. Volume 32, by Toshihiko Shiotsu, is entitled *Components of L2 Reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners*.

This latest volume investigates the linguistic and processing factors underpinning the reading comprehension performance of Japanese learners of English. It describes a comprehensive and rigorous empirical study to identify the main candidate variables that impact on reading performance and to develop appropriate research instruments to investigate these. The study explores the contribution to successful reading comprehension of factors such as syntactic knowledge, vocabulary breadth and reading speed in the second language.

Key features of the book include: an up-to-date review of the literature on the development and assessment of L1 and L2 reading ability; practical guidance on how to investigate the L2 reading construct using multiple methodologies; and fresh insights into interpreting test data and statistics, and into understanding the nature of L2 reading proficiency. This volume will be a valuable resource for academic researchers and postgraduate students interested in investigating reading comprehension performance, as well as for examination board staff concerned with the design and development of reading assessment tools. It will also be a useful reference for curriculum developers and textbook writers involved in preparing syllabuses and materials for the teaching and learning of reading.

Information on all the volumes published in the SiLT series is available at: www.CambridgeESOL.org/what-we-do/research/silt.html