BRITISH COUNCIL

TeachingEnglish

East Asia

New Directions: Assessment and Evaluation

A collection of papers

Edited by Dr Philip Powell-Davies

East Asia

New Directions: Assessment and Evaluation

A collection of papers

Edited by Dr Philip Powell-Davies

ISBN 978-0-86355-680-7

The British Council creates international opportunities for the people
of the UK and other countries and builds trust between them worldwide.

A registered charity: 209131 (England and Wales) SC037733 (Scotland).

Contents

Theories and Practices 13

THE STUDENTS' PERSPECTIVE 39

THE TEACHERS' PERSPECTIVE 97

FOREWORD

Samantha Grainger

If you are one of the many ministries of education or professionals grappling with the issues of how to effectively evaluate or assess English language learning, this collection of papers – the proceedings from the **East Asia Regional Symposium on Assessment and Evaluation** – is essential reading.

Governments across the world recognise the importance of English to their economies and societies and to the fulfilment of the personal aspirations of their citizens. There is also growing evidence and understanding of the potential for English to empower and support personal and professional development.

On a macro level, governments are managing a global recession and working populations are on the move. Migrant workers play a key role in the national economics and education of many countries in Asia. On a micro level, individuals compete for jobs in an increasingly competitive global employment market. What challenges do this present for governments? How will this affect policy and practice?

The Symposium and these Proceedings capture the insights from contributors with a wide range of backgrounds and experience, describing different contexts and tackling the issues from a variety of perspectives.

The British Council has promoted and encouraged debate on the role and the teaching and learning of English worldwide since its foundation in 1934. We remain committed to this aim today. We believe that English opens doors, creates opportunities for mobility and education, supports economic growth through international trade and helps to create new international communities – enabling people from diverse backgrounds to share ideas and opinions and promote intercultural dialogue.

Assessment and evaluation are crucial cogs in the wheel of the education system and as such need to be reliable and fit for purpose. The significance of assessment *of* and *for* learning; the critical difference in their concept, purpose and approach; the importance of teacher competency in assessment methods and test development, and the significance of robust monitoring and evaluation systems in large scale projects were key themes of the Symposium. They reinforce the understanding that all of these areas need to be tackled if we are to work towards improving the teaching and learning of English in East Asia.

I hope that these Proceedings provide helpful background and stimulus for further research in the area of assessment and evaluation. I would welcome suggestions for new research projects to deepen our collective knowledge.

Individual papers will be available on the British Council's website http://www.britishcouncil.org/accessenglish.htm. You can also find links to a wider selection of papers and research from around the world, including case studies and resources for those working in the field of ELT.

Finally I would like to express my thanks to all of the contributors who participated in the Symposium and who have written papers for this publication. I would also like to thank the Symposium Facilitator and the Editor of these Proceedings, Dr Philip Powell-Davies, for his expertise in leading the Symposium and pulling this publication together, and to Mina Patel for organising the Symposium and liaising with participants from Europe and around the region.

PREFACE

Philip Powell-Davies

The last several years have seen a great deal of change in education in East Asia particularly in the areas of the language of instruction, teacher training and curriculum reform. However, with few exceptions, there has been little evaluation of assessment and evaluation practices. There is a growing interest in this area and a number of reasons help to explain why:

- good examination results are only part of the picture and employers are increasingly indicating that workplace performance and the development of skills are just as important;
- educational institutions are coming under increasing pressure from parents and students to demonstrate learning outcomes and the quality of their teaching;
- governments, donors and heads of educational institutions are increasingly accountable to demonstrate value for money in order to report on issues of impact, sustainability, quality and relevance.

For all education institutions and decision makers at whatever level this means not only reviewing curricula and methodology but also aligning assessment and evaluation processes and instruments so that they are fit for purpose. It is clear that a one-size-fits-all approach is inappropriate and a much more subtle and eclectic approach is needed, blending formal and informal approaches and both summative and formative methods of assessing impact, learning and quality. And more broadly, systems of evaluation at school, university and in the work place need to be better integrated and linked so that positive outcomes at one level have a beneficial effect in the next, both for individuals and for the integrity of the whole education system.

The *New Directions: Assessment and Evaluation Symposium* was organised by the British Council in East Asia in July 2011 to address a number of increasingly important issues facing professionals across the spectrum of education in the region. The Symposium brought together over 100 decision makers, planners, practitioners and academics from ministries, universities, assessment agencies, schools, projects and the private sector to discuss assessment and evaluation in all its forms and complexions.

The Symposium set out to achieve a number of objectives, principle among which were:

Knowledge sharing

- Sharing knowledge about different types of assessment and evaluation in the region and further afield at primary, secondary and tertiary level
- Sharing experience and practice of assessment and evaluation systems from the region and elsewhere.

Debate and Discussion

- Creating a forum for policy dialogue to increase understanding at the strategic level of:
  - the benefits and challenges involved in developing assessment and evaluation models and systems; and
  - the challenges surrounding the effective planning and implementation of assessment and evaluation systems.

Partnership

- Working together to develop relevant, context-dependent plans of action for the improvement of assessment systems in the East Asia region.
- Identifying possible cooperation areas to develop mutually beneficial collaborative projects.

These objectives were explored in detail through a number of themes, which became increasingly rich and multifaceted as the Symposium progressed and have been developed further by several of the papers in this publication. The themes that most of the participants were drawn to were:

Assessment of/for Learning

- The relationship between formative and summative assessment
- The assessment of content being learned through English
- The challenges in developing national examination systems
- The assessment of learning.

Assessment/testing Literacies

- Training for test developers
- Teacher competence/readiness for assessing learners
- Assessment of Young Learners programmes.

Teacher Assessment

- Standards/competencies for teachers
- Assessing teaching
- Linking assessment and teachers’ professional development.

Programme Evaluation

- The development of frameworks for large scale programme evaluation
- Linking research and impact evaluation
- The modalities and implications of programme evaluation.

The contributors to these Proceedings inhabit a spectrum of cultures and represent different horizons of thought. They come from a wide range of social positions – academics, teachers, researchers, policy makers, implementers - and have been requested to present their views and expand our understanding of a range of areas because of the quality and diversity of their experience and the insights they can bring to bear. The articles presented here have at least three qualities in common. First of all, the perspectives they explore are in some ways subversive, in the sense that they turn a situation around and look at it from another angle, and assess what effect it has on people and systems. As a result of this, the articles share a second concern with the exploration of how assessment and evaluation works both at the level of the individual and at the level of the system. Several of the contributors consider the negative impact that policy decisions around assessment and evaluation can have, particularly in the context of an increasingly globalised world economy. As such, the intention of many of the contributors is not to outline a smooth, faultless landscape of well-integrated and sensitive evaluation approaches and systems, but to examine how individuals and systems can be both positively and negatively affected; how this relates to the quality of learning and teaching, and how this ultimately translates into economic opportunities in the workplace. Finally, the ideas presented in these Proceedings are in many ways radical in the sense that the contributors go to the root of the questions they are exploring, and in the process variously offer solutions and raise more issues that need to be interrogated further.

Part One deals with the relationship between formative and summative assessment and considers how a theory of test validation can be used to support an entire test development programme linked closely to a clear understanding of the value of standards. The meaning of standards is closely examined together with the relationship of standards to a developmental continuum. It is argued that when standards are interpreted in this way, their implementation provides information that enables both formative and summative purposes to be achieved.

Part Two discusses different aspects of assessment and evaluation from the perspective of the student. The case is made for specific tailored approaches to make assessment appropriate and relevant both to types of students and the social context in which learning takes place. The relationship between testing and student learning is examined in detail, in places developing the notion of learning-oriented assessment comprising appropriate assessment task design; student involvement in assessment through peer- and self-evaluation; and dialogic feedback. Implications for policy and practice are drawn out and challenges for implementation addressed. The issue of student anxiety is also examined and its effect on learning outcomes. A student learning outcomes–based accreditation model is described linked clearly to wider issues of accountability and the delivery of quality education.

Part Three explores assessment from the teachers' perspective - the development of teachers and the role of assessment in providing motivation and impetus to strengthen professional development is discussed, as well as the influence of experience in assessment processes juxtaposed with approaches adopted by novice teachers. The pervasive influence of examinations on teaching behaviour is also addressed by several contributors and suggestions proposed to address them.

It is in Part Four that we examine the assessment of specific skills areas. The articles describe the processes that contribute to listening and suggest how they might form a framework for more valid second-language tests of the skill. The importance of local context is highlighted and sheds light on where learners' problems lie. A case is made for local tests serving a 'testing for teaching' function targeted at the needs of specific groups of students. Singapore's English Language Syllabus 2010 is also showcased as a model for new directions for testing. Its emphasis on developing learners' multi-literacy and higher order thinking skills calls for a review in the function and form of assessment, which goes beyond the accustomed summative approach.

The last section of the book, Part Five, broadens out the discussion of language into the domain of the evaluation of large-scale projects and programmes. The interlocking web of variables involved in programme evaluation is developed and discussed, together with a description of an organisation's conceptual logic model to provide greater structure to project evaluation of impact with stakeholder involvement at its heart. Participatory evaluation approaches are examined together with related concepts of empowerment linked to sustainability. The role of donors, government and project stakeholders is interrogated when considering impact assessment, arguing for a clearer understanding of concepts of participation, institutionalisation and sustainability in working collaboratively with governments and communities to achieve educational impact.

We hope that readers will find these Proceedings provocative, stimulating and a spur to continue the debates raised in the papers contained here. There is a need for more research and discussion to understand the nuances of the topics explored in the book, and the British Council very much welcomes your responses to *New Directions: Assessment and Evaluation.*

THEORIES AND PRACTICES

1

Theories and practices in language testing

Barry O’Sullivan

Abstract

In this paper, I will show how a theory of test validation can be used to support an entire test development programme. Starting with a discussion of the importance of standards in any learning system, the paper goes on to consider a model of test validation, linking this to the key areas of quality and localisation. Finally, the paper outlines some of the key challenges facing the local and international test developer.

Standards

Within the context of language learning, teaching and assessment, the notion of standards has been with us for some time. Forty years ago, the Council of Europe (CofE) first began the task of describing in some detail those aspects of language which appear to be ‘criterial’ at different levels of language ability (van Ek, 1977; van Ek & Trim, 1990, 1997). This work eventually led to the publication in 2001 of the Common European Framework of Reference for Languages (CEFR) (CofE, 2001) due in no small part to the significant work of North (2000). While similar frameworks were developed elsewhere (e.g. in Canada and Australia), the CEFR was to quickly dominate the scene initially across Europe and later as far afield as Asia and South America.

The CEFR has been criticised for its lack of detail (e.g. Weir, 2005); however, it is still recognised as a valuable contributor to communication between testers and teachers as it provides a description of language that can be used across different contexts (e.g. school assessment in Ireland and business assessment in Germany) and different languages (e.g. French or Welsh). The other major change in language learning and assessment contributed to by the CEFR is in our understanding and interpretation of the concept of language level. The publication by the CofE of a manual outlining procedures for establishing the level (in relation to the CEFR) of a particular exam (CofE, 2003, 2009; Figueras et al., 2005) led to a significant change in the way in which test developers considered the centrality of level in the development and validation process. Governments across Europe began to demand that developers provide empirical evidence of any claimed link between their test and the CEFR. However, the manual and many of the studies that were conducted based on the procedures outlined in the manual were subjected to criticism due to their lack of any explicit theoretical basis (O'Sullivan, 2009). This criticism has focused in particular on the area of validation.

Validity & Validation

Traditional views of validity identify three areas on which test developers were expected to focus in order to establish evidence of the validity of any inferences that could be drawn based on performance on their test. These three areas were:

- **construct** - the underlying trait or ability being tested
- **content** - the relevance and coverage of the content of the test with regard to the construct, and
- **criterion** - how similar the results of this test are to the results of a test which focuses on the same construct.

Over the years, this view of validity has changed radically, thanks mainly to the work of Messick (e.g. 1975, 1980, 1989). Messick argued that for the inferences that are claimed based on candidate performance on a test to be seen as valid then the developer would be expected to present what he called an integrated and substantive argument based on evidence from a range of areas (mainly related to the three traditional areas described above). This view of validity (known as the unitary approach) has come to dominate assessment over the years.

Messick's contribution to the area of validity also included the notion of consequence. Consequence, in Messick's view, referred to the expectation that test developers should take into account the intended as well as the unintended consequences (or impact) of the tests they develop. This idea quickly gained support across the academic assessment community and resulted in the concept of *consequential validity*. Messick himself was not happy with this concept (McNamara, 2006), and it certainly does not fit philosophically with his conceptualisation of validity as being based on a multi-source, evidence-based argument rather than the idea of a number of different *validities*. However, the concept of consequential validity, with its focus on test impact, social value implications and ethics, prevailed and by the end of the first decade of the 21st Century had come to occupy a position apparently impervious to challenge. Theorists such as Kane (1992) and Mislevy et al. (2002, 2003) contributed to the validity debate with detailed contributions as to the nature of the evidential argument (in the case of the former) and the linking of the test development process to an underlying model of validity (in the case of the latter). The lack of practical application of these models has been highlighted by O'Sullivan & Weir (2011) while the whole concept of consequence (as envisaged by Messick) was questioned by Cizek (2011), who argued that the consequence of using a test for a particular purpose must be the responsibility of the user rather than the developer.

The most significant contribution to the practical application of validity theory to emerge in recent years has been that of Weir (2005), whose socio-cognitive frameworks have had a major influence on test development and validation. The approach outlined by Weir (2005) has recently been updated in terms of its application to test development (Shaw & Weir, 2007; Khalifa & Weir, 2009), test validation practice (O'Sullivan 2009) and test validation theory (O'Sullivan, 2011b; O'Sullivan & Weir, 2011). The approach is summarised in Figure 1 and discussed below.

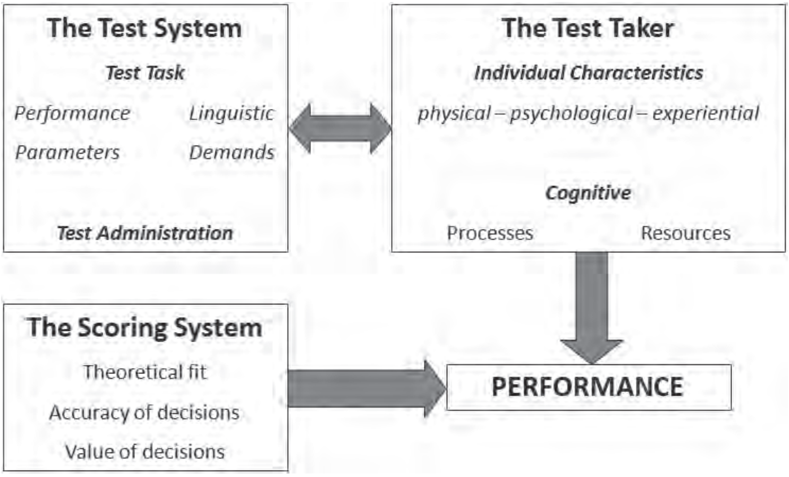


Figure 1 - A re-conceptualisation of Weir's socio-cognitive framework (from O'Sullivan, 2011b).

The main focus of the approach lies with the test taker. O'Sullivan (2000) argued that we should take into consideration three sets of test taker characteristics. These are:

Physical	This includes variables such as gender, age, and short-term ailments (such as cold, toothache etc.) and longer-term disabilities (e.g. dyslexia, limited hearing or sight etc.).
Psychological	Refers to memory, personality, cognitive style, affective schemata, concentration, motivation and emotional state. It also links to the cognitive processes and resources described below.
Experiential	This can include education as well as experience of the examination and factors such as residence in the target language country.

At the same time, the developer must take into consideration aspects of the cognitive domain as they relate to the test. These are very important as they will contribute to the operational definition of the trait or ability being tested and will also ensure that the type of processing engaged in by the test taker will reflect that of the target domain:

Processes	This refers to the cognitive and meta-cognitive processing engaged in by the test taker when responding to test tasks and items.
Resources	This relates to the test taker's knowledge of the test content and to the person's language ability.

When it comes to the test task, there are a number of parameters that should be taken into account, these will relate to various aspects of the task itself and to the administration of the test, which should at all times aim at fairness and equality for all test takers. The task related parameters include:

Performance Parameters These include things such as timing (time allowed overall, for planning etc.), item/task score weighting, knowledge of how performance will be scored and so on.

Linguistic Demands This can refer to the language of the input as well as to the language of the expected output. In tests of writing or speaking this can also refer to the audience or interlocutor.

Test administration parameters should include:

Delivery The platform (computer, live, pen and paper) should be appropriate to the construct.

Security The administrative systems that are there to ensure the security of the entire delivery process.

Physical Organisation This refers to room setup, seating arrangements, etc.

Uniformity The rules and regulations which ensure that all administrations of the test are the same for all candidates.

Finally, the test developer should consider the system that is set in place to convert test performance into a meaningful score or grade. The elements referred to in the model are:

Theoretical Fit The way test performance is assessed should fit with the conceptualisation of the ability being tested. This goes beyond a key or rating scale and refers to everything from rater and examiner selection, to training, monitoring and analysis of rating behaviour.

Accuracy of Decisions Includes the traditional view of reliability, though in reality is much broader as it should relate also to all aspects of the psychometric functioning of a test.

Value of Decisions This relates to criterion-related evidence (e.g. comparison with measures such as teacher estimates, other test scores or to performance standards such as the CEFR).

Linking Tests to Standards

The CofE Manual (2009) suggested a four stage process through which test developers should go to establish empirical evidence of a link between the important cut-scores in their tests and the CEFR. The four stages were:

Familiarisation Ensuring that the personnel who are engaged in the linking process are all familiar with the relevant skills and levels as described in the CEFR.

Specification Completing a series of checklists, designed to establish the quality of the development process.

Standardisation Conduct standard-setting events in which a panel of experts first clarify a definition of the minimally acceptable person at a particular level (i.e. the person right on the borderline between a pass and a fail, but just about on the pass side of that border) and then use this definition to establish the cut-score which is later used to identify the pass/fail boundary.

Validation Gather and present evidence of the validity of the inferences to be drawn from performance on a test – in the manual, the primary focus is on the psychometric qualities of the test.

This approach has been criticised by O'Sullivan (2009) for its implied linearity and for its failure to recognise the importance of basing any such linking project on a clearly described underlying theory of test validation. An alternative approach has been suggested by O'Sullivan (2009) and is summarised in Figure 2.

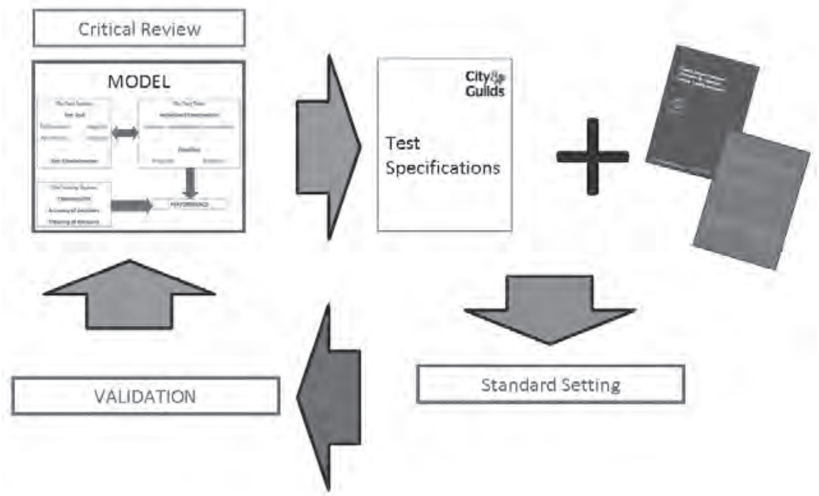


Figure 2 - The Process of Linking a Test to a Set of Standards (here the CEFR).

This process is now seen as:

Critical Review Review by an expert panel (to include individuals not connected to the test developer) which reviews all aspects of the test to ensure that it meets the level of quality required for consideration for linking (there is no point in starting any linking project without this stage - the project will break down quickly as the panels go through the specification and standardisation stages).

Familiarisation Ensuring that the personnel who are engaged in the linking process are all familiar with the relevant skills and levels as described in the CEFR. This process is repeated at all stages of the linking project (as suggested in the original CofE manual).

Specification Completing a working specification of the test. This should be based on an underlying model of validation (such as Weir 2005). Then a series of quality assurance checklists should be completed which take into account all aspects of test level and overall quality.

Standardisation This element of the process should remain unchanged – though it is recommended that exemplar tasks for the target CEFR level be included for comparison purposes.

Validation Gather and present evidence of the validity of the inferences to be drawn from performance on a test. This is a formal validity argument and should present evidence from those sources identified in the underlying model (typically those areas referred to by Messick).

Localisation Localisation is the practical implication of taking the test taker into account at all stages of the development process. For this to happen, the developer must identify any factors which can impact on test performance. By doing this we can facilitate the individualisation of assessment by identifying the key individual characteristics of members of a particular population (i.e. those identified in Figure 1 above) when developing tests for use with that population. In this way the developer is taking into consideration the consequences of all decisions make during the development process.

One important implication of localisation is the recognition of the importance of test context (which of course includes the test taker, and also the social and domain-related context) on test development. Locally appropriate tests are therefore those which are designed for use in a specific language domain. One essential element of local appropriateness is how the assessment or test ‘fits’ into the learning system. I have argued in the past that assessment must form a cohesive part of an integrated language system, in which the curriculum, the delivery of the curriculum (including teacher selection & training; teaching materials; physical setup of learning environment) and the assessment system (summative and formative) should all be based on a single underlying philosophy of language and learning.

When tests that have been created at the local level, for example in schools and universities, are compared with those developed by major international developers they are invariably considered as being inferior in a number of ways. One of these is the published internal consistency estimates which for international tests are generally very high and often for local tests are relatively low. Local tests are regularly rejected in favour of international tests on this basis, despite the fact that the variables that affect these estimates include the population’s size and range of ability. So, we are not actually comparing like with like. Ideally, international tests should publish reliability estimates for the population for which a comparison is being made.

In reality, the major differences between local and international tests are in the area of quality not so much of content relevance of coverage, but in the areas of test presentation, item and task development and administration systems. Therefore, the biggest challenge to locally developed tests lies in their ability to match internationally recognised tests in these areas. Of course, these same internationally recognised tests typically have one significant weakness which limits their value (and validity) for use in many specific domains. The weakness I refer to is the fact that they are not addressed to the population in the specific domain and are often unsuitable in terms of content, level and cultural appropriateness.

Challenges

To develop valid tests therefore, the local and international developer must face up to a number of challenges. These challenges include, test taker definition; domain language definition; test tasks designed to reflect the above; and a scoring system designed to reflect the above.

Domain test taker definition
This will typically entail identifying the test population using the elements identified in Figure 1. For example, if we are focusing on young learners, it is important to consider the appropriateness of the test tasks for use with the target group.

Domain language definition
This work is linked closely to the creation of a learning curriculum as well as accurately describing the language of the domain. We often look to educational standards to help us understand what language a learner will need to know at a particular point in their development. Local expertise will allow the test (as well as the curriculum and materials) developer to identify the level of ability expected of each language element. Together, these combine to allow for the creation of a detailed matrix of language needs and goals. This results in a locally appropriate definition of the underlying construct.

Tasks designed to reflect the above
When the domain language has been defined and the test developer has identified the level of ability expected of each language element of the domain and the relevant variables associated with the test takers have been identified, then the developer can turn to the test tasks. It is important that these tasks reflect the domain and the test takers and that they elicit the type of language and cognitive behaviour typical of the domain.

A scoring system designed to reflect the above
At the same time, the developer must consider how the test performance is to be scored. This goes well beyond simple (or even complex) estimates of internal consistency or inter/intra rater reliability to include all aspects of the scoring system (e.g. rater selection, training and monitoring; rating scales and/or answer keys; test item or task analysis; grading system; and score interpretation or value). The fact that the developer will have included in the domain definition a detailed description of both the language to be tested and the level of ability required of the target population will significantly support this set of processes.

References

Cizek, G. J. (2011). ***Reconceptualising Validity and the Place of Consequences.*** Paper Presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, April 2011.

Council of Europe. 2001. The Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press. (Also available at http://www.culture2.coe.int/portfolio/documents_intro/common_framework.html.

____ (2003). ***Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEF)***. Manual: Preliminary Pilot Version. DGIV/EDU/LANG 2003, 5. Strasbourg: Language Policy Division.

____ (2009). ***Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A Manual***, Strasburg: Language Policy Division. http://www.coe.int/T/DG4/Linguistic/Manuel1_EN.asp.

Figueras, N., North, B., Takala, S., Verhelst, N. & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: a manual. ***Language Testing, 22, pp. 261-279.***

Kane, M T (1992). An argument-based approach to validity, ***Psychological Bulletin 112 (3), pp. 527–35.***

Khalifa, H. & Weir, C. J. (2009). Examining Reading: Research and practice in assessing second language reading. ***Studies in Language Testing 29***. Cambridge: Cambridge University Press and Cambridge ESOL.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. ***Language Assessment Quarterly, 3, pp. 31-51.***

Messick, S. (1975). The standard program: Meaning and values in measurement and evaluation. ***American Psychologist, 30, pp. 955-966.***

____ (1980). Test validity and the ethics of assessment. ***American Psychologist, 35, pp. 1012-1027.***

____ (1989). Validity. In R. L. Linn (Ed.), ***Educational Measurement (3rd edition)*** London, NY: McMillan.

Mislevy, R. J., Steinberg, L. S. & Almond, R. G. (2002). Design and analysis in task-based language assessment. ***Language Testing 19 (4), pp. 477–96.***

____ (2003). On the structure of educational assessments, ***Measurement: Interdisciplinary Research and Perspectives 1 (1), pp. 3–62.***

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

O'Sullivan, B. (2000). Exploring Gender and Oral Proficiency Interview Performance. ***System, 28, pp. 373-386.***

____ (2009). ***City & Guilds Communicator Level IESOL Examination (B2) CEFR Linking Project Case Study Report.*** City & Guilds Research Report. Accessed from: http://www.cityandguilds.com/documents/ind_general_learning_esol/CG_Communicator_Report_BOS.pdf

____ (2011a). Introduction. In B. O'Sullivan (Ed.) ***Language Testing; theories and practices.*** Oxford: Palgrave Macmillan.

____ (2011b). The City & Guilds Communicator Examination linking project: a brief overview with reflections on the process. In W. Martyniuk (Ed.) ***Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's Draft Manual.*** Cambridge: Cambridge University Press.

O'Sullivan, B. & Weir, C. J. (2011). Language Testing and Validation. In B. O'Sullivan (Ed.) ***Language Testing; theories and practices.*** Oxford: Palgrave Macmillan.

Shaw, S & Weir, C. J. (2007). Examining Writing in a Second Language. ***Studies in Language Testing 26.*** Cambridge: Cambridge University Press and Cambridge ESOL.

van Ek, J. A. (1977). ***Threshold Level for Modern Language Levels in Schools.*** London: Longman.

van Ek, J. A. and Trim, J. (1991). ***Waystage 1990.*** Cambridge: Cambridge University Press.

van Ek, J. A. and Trim, J. (1997). ***Vantage Level.*** Strasbourg: Council of Europe.

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

2

A standards perspective on the relationship between formative and summative assessment

Jim Tognolini and Gordon Stanley

Abstract

Most education systems have moved to a standards referenced system of reporting educational outcomes. This type of reporting requires judges (teachers) to reference assessments to predetermined standards of performance. In this type of reporting a grade is only given to those students who have demonstrated the criteria for the grade. Such reporting makes sense when we want to interpret student outcomes in terms of explicit standards. In this paper we discuss the meaning of standards; the relationship of standards to a developmental continuum; and how, when standards are interpreted in this way, their implementation provides information that enables both formative and summative purposes to be achieved.

The meaning of assessment

Most recently we have started to define assessment as *involving professional judgment based upon an image formed by the collection of information about student performance.* In education, students are generally assessed for the purpose of improving their learning, and monitoring and certificating their performance or achievement.

Teachers collect information about student performance (assessment) in numerous ways. These have been summarised along a continuum of assessment methods (Figure 1) that range from 'less formal or unstructured methods' to 'more formal or highly structured methods' of collecting information.

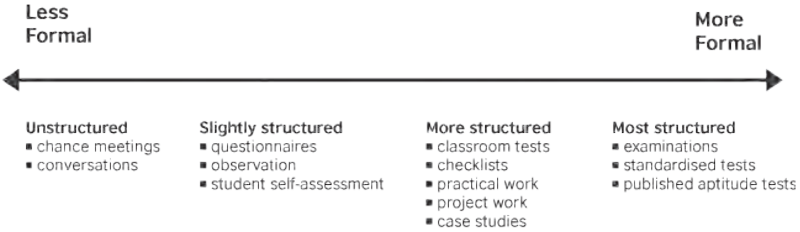


Figure 1 - Methods for collecting information on student performance.

At the highly structured end there are examinations, published tests and tests such as national and state-based testing programmes. These are highly structured in that the conditions of administration are tightly controlled and standardised; the tests have been through rigorous test construction processes, and so on.

Classroom tests, checklists, practical work, project work, etc. are also methods for collecting information about students. They are not as formal in their structure but they provide information that is just as pertinent and relevant about a student as the more highly structured means of collecting information and they happen much more often.

When teachers ask a question in class they are assessing. When they observe what is going on in class they are assessing. When assessment is thought about in this way it seems obvious that there is no division between assessment and teaching. They are one and the same.

Sometimes teachers tend to play down the importance of the assessment information that emerges from the less formal end of the assessment continuum because it does not generally culminate in marks that can be recorded in marks books. However, when they do give a more structured assessment they already have an expectation of how the students will perform built up from the less formal (formative assessment) that takes place continuously; and, they validate the “summative” performance against the expectation or image of the student that has already been built up from the wide range of assessment activities that take place constantly in the classroom situation. The latest piece of information contributes to the image.

Generally the information that emerges from the test, standardised test or examination is consistent with the image. Sometimes it is not and the teacher then asks the question, “Why not?” There are many students who perform well in classroom activities and yet perform poorly in the examination; such atypical performance is of interest to teachers. It could be that the student has really improved and there is a need to adjust the image. Alternatively, it could be that the result may be due to other reasons that would not warrant the substantive change in the image.

In summary, therefore, teachers use assessments to form an image of what students know and can do. As more and more information becomes available from a variety of assessment sources, it is added to the image. The various forms of assessment are assessing the same material from different, but interrelated perspectives. Consequently the “fairest image” emerges when teachers use a range of assessment techniques and assimilate the information from the multiple sources using their professional judgment.

Types of standards

Commonly a distinction is made between curriculum standards and performance standards. **Curriculum standards** are defined as “the knowledge, skills and understanding expected to be learned by students as a result of studying a course”, while **performance standards** are “the levels of achievement of the knowledge, skills and understanding.”

Some terms synonymous with **curriculum standards** are:

- Syllabus Standards
- Content Standards
- Grade Level Standards
- Core Standards
- Outcomes
- Competency Standards (VET and Professional).

Those which tend to be used interchangeably with **performance standards** include:

- Achievement standards
- Benchmark Standards
- Proficiency Standards
- Reporting Standards
- Accountability/Target Standards
- Performance Indicators.

Standards referenced assessment

Traditionally marks evolving from assessments have been given meaning by referencing them to norms (norm-referencing). In the 70s and 80s systems moved towards referencing student performance against criteria (criterion-referencing). More recently, significant numbers of education systems around the world have introduced a different way to reference achievement. It builds upon criterion-referencing, but instead of referencing achievement to the myriad of behaviours that comprise an examination, course or subject, the achievement is referenced to pre-determined standards of performance. It is referred to as standards-referencing.

The following characteristics are required of well-grounded standards referenced systems:

- Standards should describe performance expectations and proficiency levels in the context of a clear conceptual framework, and are built on sound models of student learning (developmental continuum).
- Standards should be clear, detailed, and complete; reasonable in scope; and both rigorous and well-grounded in the knowledge domain.
- Standards must be elaborated so that curriculum, teaching and assessment are aligned.
- Standards are derived from the curriculum and not developed independently from mandated curriculum requirements.

The value of standards referenced assessment systems is that they:

- enable the performance of cohorts of students to be monitored against pre-determined standards;
- empower students in the teaching and learning process;
- provide all students with a meaningful record of their achievements;
- provide a mechanism to recognise and reward prior achievement at school within broad-based qualifications frameworks; and,
- provide a mechanism to bring together curriculum, pedagogy and assessment in a way that has not been possible in the past.

The developmental continuum

One of the main ideas that has emerged in relatively recent times is the notion of developmental assessment. This is the process of monitoring a student's progress in a subject so that decisions can be made about how to improve learning for the student. Developmental assessment shifts the focus of attention in assessment from comparing one individual to another, towards one of monitoring student progress. The key feature of developmental assessment is that the students' progress or growth in the subject is monitored along a linear continuum that is referred to as a developmental continuum (See Figure 2).

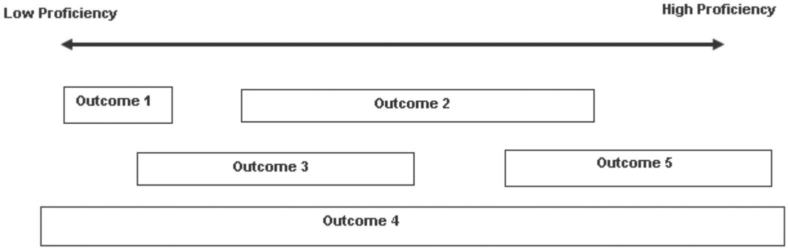


Figure 2 - Schematic representation of a developmental continuum.

The monitoring of student growth along a developmental continuum requires that the continuum be defined. Many countries have now defined continua for the various subjects in terms of learning outcomes. These outcomes typically describe what students know and can do at different stages along the continuum. These outcomes are usually contained in syllabus documents or frameworks and provide the basis for the development of the teaching and learning sequence and activity (including assessment) within the subject.

It can be seen from Figure 2 that some of the learning outcomes extend across the whole continuum (e.g. reading for meaning) whereas others are relatively less extensive. The further the outcome extends across the continuum, the more demanding it is for the students and the more of knowledge, skill and understanding of the subject is required to demonstrate achievement of the outcome.

To progress along the continuum students have to become more proficient in the subject. Similarly learning outcomes that are further along the continuum are more demanding for the student. They require more of the "property", "trait" or "thing" that defines the subject to be able to demonstrate proficiency. The whole idea is based upon growth.

Generally the developmental continua are partitioned into levels, stages, bands or grades (see Figure 3). The grades have descriptors (grade related descriptors) that try to capture the skills, understanding and knowledge that students have at different stages along the developmental continuum for the subject. These represent broad descriptions of standards and teachers in schools and examiners are able to locate students along these continua by comparing their "images" of students to these broad standards and using their professional judgment to say, on balance, that the student is located at "Grade D" or "Grade A" at this stage of their learning. Just as importantly, students can also locate themselves along this continuum by judging their own performance and work out what they have to do to go from a lower grade to a higher grade along the continuum. The continuum is cumulative in that what is required for a Grade C is everything that is required for a Grade D and Grade E plus the extra for a Grade C. Similarly, a Grade A includes everything that is required for all grades up to A, plus the extra segment unique to Grade A.

In order for students to demonstrate where they are along the continuum, they must be given the opportunity to demonstrate what they know and can do in relation to the outcomes of the subject. Tasks or items that examiners and teachers set provide this opportunity for the students to demonstrate what they are capable of doing.

In the case of formal (summative) assessments, such as public examinations and standardised tests, the examiners or test constructors must write items to match the student learning outcomes that are in the syllabus documents so that the results can be interpreted in terms of the same developmental continuum that is being used by the teacher in the classroom. In this way the results should be providing one more piece of information about the location of the student and should supplement the teaching/learning process that is going on in the classroom. Figure 3 shows items and a student along the continuum.

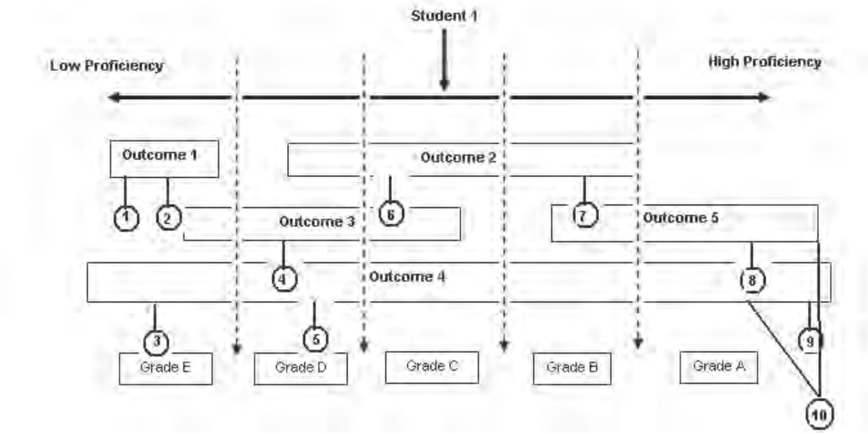


Figure 3 - Developmental continuum with items, students and grades.

Item numbers in Figure 3 are denoted by the numbers in the circles. It can be seen therefore that item 1 in the test is assessing Outcome 1 and is relatively easy (because it is located towards the left on the continuum); item 2 is further to the right of item 1 on the continuum so it is demanding more of the student and hence it is harder than item 1 and still measures Outcome 1. Item 3 is a bit harder than item 1, not as hard as item 2 and is measuring outcome 4. Item 10 is a bit harder (hence it is more demanding of the students) than all the other items and actually measures two outcomes: outcome 4 and 5.

Similarly it can be seen from Figure 3 that Student 1 is located within Grade C on this particular continuum. Because this student is located at that point along the continuum it could be expected that the student would get the items that are less demanding (easier - that is, they are located to the left of the student) correct, and the more demanding items (harder - that is, they are located to the right of the student) incorrect. Of course students do not always behave in such an orderly fashion. They will probably get some easier items incorrect and some of the harder items correct. This is useful diagnostic information for both the student and the teacher.

In the classroom situation teachers can make an “on balance judgment” about the location of the student on the continuum.

In the case of the formal assessments (assessment of learning), the number of marks that the student gets on the examination locates the student along the continuum: the more marks students get on the assessment, the further they are located along the continuum.

Setting assessment tasks in a standards-referenced system

In a standards referenced system, tasks (items or questions) should be set in a way that provides evidence of where the students are located along the developmental continuum. Some basic task development hints would include ensuring that:

- the items and tasks match the content standards (outcomes) articulated in the syllabus;
- the items, and tasks that are developed enable students at different stages in their learning to demonstrate what they actually know and can do; and,
- a range of different tasks are used to generate a reliable and valid estimate of the student’s location along the developmental continuum.

Standards referencing serving both formative and summative purposes of assessment

In a standards referenced system the actual marks should represent locations along a detailed continuum. They have meaning. Consequently when constructing tasks and items the teachers and test developers should keep in mind the basic tenets required of the continuum; that is, more marks should imply more of the property being assessed. Therefore, when constructing tasks the marking rubrics (and options in the case of multiple choice items) should reflect the theory. If this is done then every response can be interpreted in terms of location and should give an indication of what needs to be done to improve. For example, consider the following task:

$$2\tfrac{3}{8} + 2\tfrac{4}{7} = ?$$

The following marking rubric demonstrates how student responses can be recorded and used to inform learning.

4 marks Adds simple fractions correctly. Demonstrates correctly that only like fractions (same denominators) can be added. Obtains correct answer.

e.g. $2\frac{3}{8} + 4\frac{4}{7} = 6(\frac{21}{56} + \frac{32}{56}) = 6(\frac{53}{56})$

3 marks Adds simple fractions correctly. Demonstrates correctly that only like fractions (same denominators) can be added. Makes an error in the final step.

e.g. $2\frac{3}{8} + 4\frac{4}{7} = 7(\frac{21}{56} + \frac{32}{56}) = 7(\frac{53}{56})$

2 marks Adds simple fractions correctly. Demonstrates correctly that only like fractions (same denominators) can be added. Does not demonstrate an understanding of how to convert to like fractions and makes serious errors.

e.g. $2\frac{3}{8} + 4\frac{4}{7} = 6(\frac{3}{56} + \frac{4}{56}) = 6(\frac{7}{56})$

1 marks Does not add simple fractions correctly. Does not demonstrate an awareness that fractions have to be alike before adding. May convert mixed to improper fractions correctly but shows know understanding as to why this should be done. Shows some elementary understanding of fractions.

e.g. $2\frac{3}{8} + 4\frac{4}{7} = (\frac{19}{8} + \frac{32}{7}) = (\frac{19}{8} \text{ x } \frac{32}{7}) = (\frac{608}{56})$

0 marks Does not add simple fractions correctly. Basically shows no understanding at all of what to do with fractions.

e.g. $2\frac{3}{8} + 4\frac{4}{7} = (\frac{19}{8} + \frac{32}{7}) = (\frac{51}{15})$

It can be seen that the item could be given as a multiple choice item and there is significant information in the response to help students improve their learning.

Conclusion

The authors are firmly of the view that if assessment tasks are constructed (formative or summative) with a developmental continuum in mind then it is possible that the information can be used to help improve learning irrespective of the type of task. There are many other examples that could be used. However, all have the same basic features as the example used above.

3

Formative assessment in primary English classrooms in Vietnam

Pham Lan Anh

Abstract

Following the emergence of formative assessment as a valuable practice in assisting learning, a number of investigations have been conducted to research whether its implementation in a particular educational context actually succeeds as it is claimed to. Starting from the hypothesis that primary teachers' assessment procedures are much influenced by a traditional assessment approach, this case study attempts to reveal the extent to which current assessment practices in several primary schools in Ha Noi, Viet Nam facilitate learning. This paper starts with an overview of the context of teaching English at primary level, examining factors affecting teachers' assessment practices, namely the policy, curriculum, the learning and teaching environment. Next, preliminary results from the investigation are analysed and discussed. Finally, this paper reaches conclusions regarding the formative elements in the researched classrooms.

1. The Context of Teaching English to Young Learners in Ha Noi

1.1. Policy and Status of English

Although English has been recognised as the first foreign language, and the Vietnamese Government clearly underlines the importance of developing English to better compete and showcase the skills and talents of its workforce in regional and global markets, English in Viet Nam is still treated as a subject for study rather than as a living language to be spoken in daily conversation. Within the framework of the Master Plan on Foreign Languages Teaching and Learning in the National Education System 2008- 2020, English teaching and learning, which is supposed to be implemented in stages, is to follow a 10 year compulsory curriculum, starting from Grade 3, with a time allocation of 4 periods of 40 minutes per week. Against this backdrop, in Ha Noi, English is still being officially taught from Grade 3 as an *optional* subject, 2 periods/week, and has not yet been included in children's achievement records. On the one hand, compared to other subjects at primary school, English is thus viewed from teachers', parents' and children's perspectives as less important and less serious. On the other hand, in practice, teachers still give tests periodically to children as a means of collecting information on learning/ progress.

In 2003, the official curriculum for primary English was approved and has been revised ever since. The latest revised curriculum is claimed to take account of the needs of young learners in primary school, which are different from the needs of older children in secondary school. As stated in the document, the principle of developing the primary English curriculum is to emphasise communicative competencies and therefore seeks to promote more communicative teaching methods, through coherent themes and topics which are meaningful and relevant to the student's world. The guiding principle also ensures that there is a recognition that primary age students are still developing cognitively. They are not able to think abstractly or to analyse the structure of languages. The teaching methods need therefore, to be child-centred, based on actions and with many opportunities to practise language skills in meaningful contexts.

Regarding assessment, after following the primary education English curriculum, children will be considered to have mastered the equivalent of Level A1 of the Common European Framework of Reference for languages. However, the curriculum does not make explicit the language ability expected of pupils, nor does it align with expected competencies or standards, which should have been included in the curriculum.

1.2. Teaching and Learning Environment

Within the scope of this paper, the teaching and learning environment here is limited to the physical environment, social situation and instructional arrangements.

The physical environment in most primary schools in urban areas of Ha Noi has been gradually improved so as to enhance teaching and learning. Classrooms are furnished with adequate desks and chairs, good lighting and with paucity of modern multimedia. However, such improvements have led to large class size, ranging from 55- 65 pupils in a space which used to accommodate 40. This problem influences English language classrooms and leads to teacher difficulties: failure to manage activities *such as mingle, group work, survey, story telling* - in which children are supposed to move around the classroom. This, to a greater or lesser extent, affects the effectiveness of an English lesson.

The physical environment, therefore, implicitly influences the nature of classroom interactions. Moreover, techniques of cooperative and communicative learning require children to socialise comfortably with one another, but fixed seating arrangements hinder these. Scaffolding to meet the needs of the individual learner, as a result, is also difficult to implement.

Therefore, In terms of social and instructional arrangements, teaching English in primary schools generally has not met the criteria for contemporary teaching/ learning approaches. As mentioned in **1.1** the primary curriculum is not directive in a sense that it does not give transparent and explicit guidelines for schools and teachers to adopt appropriate teaching methods and plan their teaching and learning process accordingly. As Moon (2005) asserts:

> *"many primary teachers are using a fairly formal approach to teaching children which could be seen as more suitable for secondary pupils and adults than children … Observations indicated that some teachers using English had little idea how to support children's understanding and adjust their language to children's level. …There was a heavy emphasis on the form of the language, focusing fairly explicitly on grammatical points and trying to explain them."*

2. Current mode of Teacher Assessment

The ideas in this section derive mainly from focused classroom observations and informal interviews conducted in 7 primary classrooms in Ha Noi, with a total of 29 class hours. However, in order to triangulate and ensure validity, the findings are also supported by preliminary results of a questionnaire on teachers' perceptions and assessment practices.

The observations have been conducted over a two-year period (2009-2011) in 3 primary schools in Ha Noi. Three teachers (A, B, C) were observed for 4 class hours each, whereas one teacher (D) has been observed in two stages, with the first stage being similar to that of the other three teachers, the second stage being intensive with observations during her lessons for 15 weeks for a total of 15 class hours. These four teachers were all teaching children of Grade 3 State schools, aged from 8 to 10, who are supposed to officially start learning English. The following are common features and trends from the focused observations, informal interviews and analysis of the questionnaire. Some of the features are listed in Themes 1- 4:

Theme 1: Structure of a common English lesson

- A PPP structure (presentation, practice, production)
- Role of teacher as knowledge transmitter, role of children as recipients
- Whole class teaching and questioning.

Theme 2: Procedures of a common lesson of English

- Start with a warm up recapping the content previously taught
- Introduce new vocabulary for the new input
- Set the context for the language pattern
- Check understanding on form, pronunciation and meaning (of the language pattern)
- Make sure pupils grasp these new ideas by a question-and-answer as whole class teaching
- Practise examples by working first as a class or group and then individually
- Call some children for classwork or boardwork or games
- Give feedback on classwork or boardwork
- At the end of the lesson, look back and review the new learning and link it to previous skills and knowledge acquired
- Set homework or give worksheet for self-study
- Ask children to revisit and improve the classwork (as part of homework).

Theme 3: Common Assessment Tools

- Check understanding during presentation stage
- Observe and monitor classwork, individual work in practice/ production stages
- Give feedback (informative & evaluative) in practice/ production stages
- Assign homework in the form of worksheets or exercise book for self-study
- Assign class tests or boardwork or worksheets for marking and grading (on-going assessment)
- Give children opportunities to revisit and improve the checked work
- Sit children for end-of-year achievement test, followed by marking & grading for accountability (summative assessment)
- Report children's progress and achievement to stakeholders periodically (preferably at the beginning and end of the academic year).

Theme 4: Assessment procedures

- Planning assessment in mind
- Implementing assessment in teaching process
- Improving assessment in mind
- Most common time for formal assessment: beginning (diagnosis test) & end of academic year (achievement test)
- Most common time for informal assessment: daily: warm-up & production s tage; end of unit, mid-term, end of term 1.

As can be seen from the themes above, teachers have a tendency to underestimate the role of children as active learners who can become responsible (if supported) for their own learning. Regarding assessment, the most common form is test and worksheet. Teachers seem to test mainly knowledge of words and grammar in isolation but do not test grammar in use. There is very little use of listening tests and even less use of oral tests

Overall, the range of current assessment practices in primary English classrooms is generally summative, with a paucity of elements which might be described as formative. Annual tests and class tests and worksheets, the correction of class work and home work are the most common ways of collecting information, whereas the practices related to record keeping – to be passed on to stakeholders - are restricted to quantifiable achievement reports.

It is also common practice for primary teachers to mark and correct children's work with evaluative comments, whether the work is classwork, homework or tests. It is interesting to note, however, that the teachers observed then asked children to revisit and improve this checked work.

Two other common practices were observed: a) class discussion once the work is returned to the children; b) sending work home to receive parents' signatures to ensure that parents are informed of progress and achievement, and therefore, can provide comment/guidance to children's work at home.

3. Conclusions

Firstly, current assessment practices are very much embedded within the traditional culture of examinations and testing. This type of assessment does not take into account that children may well still be developing physically, psychologically and cognitively.

Secondly, current assessment practices have not succeeded in helping children to become aware of their own strengths/weaknesses and to learn to monitor their own/their peers' progress. As a result, children have not become involved in the learning process and there is evidence that they respond negatively to teacher feedback.

Thirdly, teachers neither regularly document the planning of assessment activities, nor do they record evidence of children learning. Instead, they merely present/ have an overall informal or anecdotal impression of children's learning. Everything appears as if *'inside the blackbox'.*

Although the range of current assessment practices in primary English classrooms is highly traditional, it may partly facilitate learning in the following areas:

- The class tests and worksheets that are designed by the teacher, based on classroom work, are non-threatening and designed to give immediate feedback to children. Children are also encouraged to assess what they have learned in particular units;
- The correction of class and home work includes comments on the children's exercise books and test papers. This might be seen as positive in the sense that children can see their weaknesses and are then encouraged to revisit and improve this checked work;
- Assigning an appropriate amount of homework at levels that match children's learning may lead to improvements in understanding and achievement via home support and reflection by the children on their work – while at home. It may also help to support positive attitudes towards learning as well as self-study habits. This is particularly true when the time allocation for English is a mere 2 class periods per week: without homework, children are likely to forget what has just been learnt.

References

Black, P., & Wiliam, D. (1998). ***Inside the black box: Raising standards through classroom assessment.*** King's College London School of Education.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). ***Working inside the Black Box: Assessment for Learning in the Classroom.***

Black & Jones (2006). Formative assessment and the learning and teaching of MFL: sharing the language learning road map with the learners, ***Language Learning Journal, Winter (4), pp. 4 -9.***

Halliwell S. (1992.) ***Teaching English in the Primary Classroom.*** Longman, Singapore.

McKay P. (2006). ***Assessing Young Language Learners.*** CUP, London.

Hayes D. (2008). ***Primary English Language Teaching in Vietnam.*** Unpublished report. Hanoi, Vietnam. British Council.

McKay P. (2006). ***Assessing Young Language Learners.*** CUP, London.

Moon J. (2005). ***An Investigation into Teaching English to Young Learners - 'Conference on English Language Teaching at Primary Levels'.*** Hanoi: Ministry of Education and Training/ British Council. Vietnam June 2-3 2005.

Rea-Dickins, P., & Gardner, S. (2000). Snares and silver bullets: Disentangling the construct of formative assessment. ***Language Testing, 17 (2)***, pp. 217–244.

Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. ***Language Testing, 18 (4), pp. 429–462.***

Torrance, H. & Pryor, J. (1998). ***Investigating Formative Assessment – teaching, learning and assessment in the classroom.*** Taylor & Francis Inc., USA.

THE STUDENTS' PERSPECTIVE

4

Assessment in young learner programmes

Sophie Ioannou-Georgiou

Abstract

As the starting age for English as a Foreign Language is officially lowered in more and more countries, education providers around the world have to face issues of accountability of their innovation while bearing in mind the particularities of early language learning (ELL) programmes, such as the specific characteristics of the young learner (YL) population and the increased importance of attitudinal and motivational issues. The author suggests that policy makers should not rely on summative exams and ignore the importance of classroom-based assessment, which can act as a significant step towards the success of YL programmes. The paper proposes a suggestion for an assessment system for YL programmes.

The growth of young learner programmes

Younger and younger learners are being introduced to foreign language learning. Many countries around the world are adopting early foreign language learning as part of their educational policy and including foreign languages as a compulsory subject from the first year of primary education (including Spain, Poland, Italy, Cyprus, Greece, Croatia, Norway, Serbia, the Czech Republic, Austria, and Malaysia) or sometimes even in pre-primary education (Spain, for example, includes foreign language in preprimary as a compulsory subject in certain areas. Nevertheless, more countries include foreign language learning in pre-primary as an optional subject (e.g. Cyprus, Finland, Poland, Slovenia, France, etc.).

There are many reasons governments are taking the decision to introduce ELL. Foreign language skills are seen as economically valuable as they can enhance trade, mobility, employability and offer an overall competitive edge in the market.

Generally, however, educational systems have always tried to prepare future citizens who are able to cope with the demands and challenges of the society they will be living in. Modern globalised societies require citizens with foreign language skills, and governments prepare their future citizens by providing them with the skills required. The European Union has, for example, specified the key competences its citizens should have so as to be able to respond to the needs of our fast-changing world (European Commission, 2007). The eight key competences proposed by the EU are competences to ensure lifelong learning and development and one of these is ‘communication in foreign languages’.

There are often strong sociocultural objectives which also influence the decision for early foreign language learning. This is usually the case when the languages of the neighbours or of minority groups are taught. Sociocultural objectives feature highly in the policy documents of the European Union (European Commission, 2005, 2008, inter alia). For the EU, which consists of 27 countries and 20 official languages, the development of linguistic tools for communication between people and the cultivation of a culture of respect of otherness and cultural diversity is vital.

The introduction of ELL is consequently an effort to respond to the challenges mentioned above. Although one has to admit that language learning can also successfully be achieved with a later start, there are a number of benefits to ELL. There are, for example, practical benefits such as more time made available for language learning which in turn can allow more languages to be added to the school curriculum. There are also benefits of sociocultural nature, such as the early development of an openness towards and appreciation of cultural diversity. Finally, there are benefits which directly involve language learning such as the development of positive attitudes towards languages and language learning. Attitudes can be formed from a very young age and have a tremendous impact in the future development of learners. Often, negative attitudes to languages can be already in place when language learning is scheduled to start. This is sometimes the case with boys whose lack of willingness to learn foreign languages (Clark, 1998; Kissau, 2006) has been partly attributed to negative attitudes towards languages (Loulidi, 1990; Loulidi, 1200; Kissau, 2006). Similarly, early formation of negative attitudes to new, ‘alien’ cultures and an early onset of xenophobia can also take place and thus negatively influence language learning (MacNaughton, 2009).

Are summative exams the best way to assess young learner programmes?

ELL programmes are, nevertheless, an innovation for educational systems around the world, and as such they are sometimes surrounded by debate and resistance. Such reactions place educational systems under pressure to prove the effectiveness and general success of their innovation. An initial response is sometimes to turn to external or internal summative exams. This decision is often affected by two main factors: parents and exam boards. Parents often ask to see proof of effectiveness in the traditional form they are accustomed to: test results. Exam boards, on the other hand, are eager to participate in this large, emerging market and have provided a number of tests which make every effort to reflect the main teaching approaches implemented in teaching YLs.

There are undeniably benefits to using external (or internal) summative tests for establishing accountability. Turning to external tests, for example, may indeed be an easy solution. They are already available, are usually well-respected by parents and teachers, are perceived to be objective and have a large mechanism to support them (design, implementation, rating, reporting, etc.).

Unfortunately, there are a range of issues which argue against external summative exams for YLs. A summative exam, no matter how well-meaning, is a source of anxiety for young students. Exams are also a source of anxiety for teachers whose livelihood sometimes depends on their students’ exam results. Teacher anxiety is, however, unconsciously transmitted to YLs who often find that their fun and exciting lessons have turned into stressful sessions of never-ending practice of exam items. If teachers see test results as evidence of their effectiveness, the unfortunate result will be teaching to the test.

Although the above negative effects of summative testing are relevant to all learners, they are more pronounced and have intensified results in YLs. Attitudes in YLs are fragile as they are at an early formative stage and can easily change. It is very common to hear stories of children who enjoyed their language lessons but later developed negative attitudes due to test-related anxiety, loss of interest after many boring lessons of exam preparation or negative test results.

It is important, before a decision on assessment of YL programmes is taken, to review the aims and objectives of these programmes. Although many language schools may focus on purely linguistic goals, this may not be what is best for such YLs at a formative stage in their development. When it comes to state educational systems, this is a non-negotiable issue. Language education is part of a child's overall education and a range of aims and objectives, beyond the purely linguistic goals, need to be taken into account.

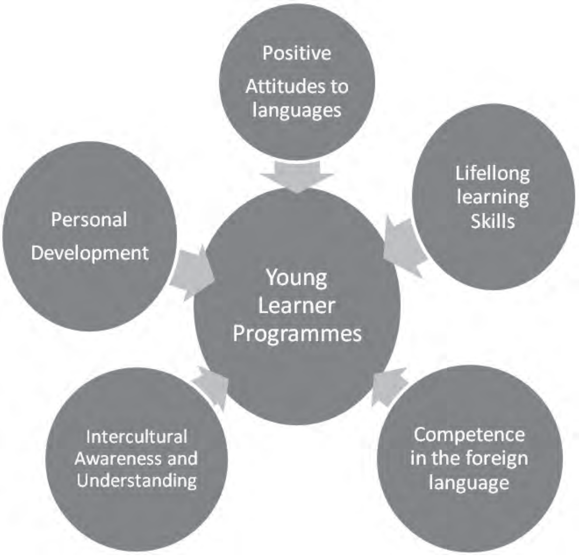

Positive Attitudes to languages
Lifelong learning Skills
Personal Development
Young Learner Programmes
Intercultural Awareness and Understanding
Competence in the foreign language

Figure 1 - Focus areas in YL programmes.

Figure 1 shows the range of focus areas which come into play in a YL programme. The importance of adding clear objectives towards the development of positive attitudes and cultivating intercultural awareness and understanding has already been discussed. Other areas of focus are personal development and lifelong learning skills.

All the areas are interrelated. Personal development, for example, can involve the development of social skills and creativity; both important for language learning. Lack of social skills will, for instance, interfere with the ability to interact with others and engage in cooperative tasks. Creativity affects use of language and expression. In the same way, learning skills can involve the development of compensation strategies for communicative problems or the development of autonomy and self-reflection skills.

All of these objectives are especially important during the formative years of a child. National education programmes clearly state their importance and language programmes need to comply with the general aims and objectives of the national curriculum.

An example of a national ELL programme's aims is presented below:

> *"The main aim of the foreign language lesson is for students to develop positive attitudes towards English and foreign languages in general and acquire a basic Intercultural awareness and intercultural skills as well as general communicative skills so that they will use English in a creative manner to achieve communication in various everyday situations."*
> *(Cyprus Ministry of Education and Culture, 2010)*

The aims of the above ELL programme are reflected in the curriculum guidelines and the syllabus which is described in terms of language, language learning skills and intercultural aspects. Furthermore, performance indicators are specified not only in terms of language skills but also in terms of learning strategy development and intercultural development (Cyprus Ministry of Education and Culture, ibid).

It is, therefore, argued that a careful consideration of what a YL programme entails should make absolutely clear the unsuitability of summative exams. The inability of a summative exam to assess all areas of a YL language programme is a very serious disadvantage. Indeed, the author would dare suggest that assessing a national ELL programme through summative exams would be equal to cheating the society for whom the national curriculum was prepared. That is to say, the curriculum would, in essence, be narrowed down to what could be assessed through a summative test. Consequently, aspects such as attitudes towards language learning or intercultural would not be emphasised during teaching.

An alternative suggestion to summative exams

In line with the above arguments, assessment of YL programmes should cater to the well-rounded development of learners, while avoiding high anxiety levels and the development of classrooms which lack enthusiasm and creativity.

The answer might lie in formative, classroom-based assessment which is arguably more suited to YL programmes, where there is no end result in the form of securing a place at a prestigious school or acquiring an important certificate. YL programmes are the beginning of a long process of language learning and expectations should be realistic and relevant to the age of the learners. Formative assessment can, therefore, be the main form of assessment helping children and teachers to improve the learning process, keeping assessment directly linked to what happens in the classroom and promoting self-reflection and self-assessment.

Formative, classroom-based assessment can take many forms: observation, project work, portfolios, conferencing, game-like activities which do not involve pencil or paper, classroom tests, and more. The large variety offers room for manoeuvre and thus teachers can become active agents preparing assessment tasks which cater to their students' individual needs and to the needs of the specific programme they follow. The variety of assessment tasks allows for assessment of all areas of the curriculum, thus avoiding the need to narrow down to what can only be assessed in an exam. Teachers can assess attitudinal progress based on classroom observation and student conferencing while development in learning strategies can be assessed through group tasks or project work.

Nonetheless, formative assessment is not yet generally accepted as a process which can provide accountability and satisfy stakeholders with proof of successful learning outcomes. Arguments against this lie mainly in the fact that classroom-based assessment is not considered to be objective and standardised. Parents and other stakeholders may also not be able to accept it or understand it, as it is very different to the kind of numerical, norm-referenced assessment they have been used to.

When deciding what is best for YLs and for the education systems aimed at by societies around the world, the answer leans towards formative, classroom-based assessment. When accountability is discussed, it needs to be viewed as accountability towards students, parents and the society in general. Surrendering the curriculum to external exam boards with ready-made tests or narrowing down the curriculum to what can be assessed through summative exams leaves much to be desired.

A model for assessment of young learner programmes

An assessment system for YL programmes needs to satisfy a range of characteristics, including:

a) The assessment system should be congruent with the curriculum. This means that the assessment system addresses all the areas of educational focus and not only linguistic targets. As most teachers are naturally influenced by assessment educational goals that are not assessed will most probably not be adequately promoted (Smith, 1991).

b) The assessment tasks employed and the overall attitude towards assessment should be non-threatening. This will allow for the anxiety-free pedagogy promoted for YLs to be fully implemented. Furthermore, YLs tend to be motivated through intrinsic motivation by what goes on during their language lesson. If the lesson is not enjoyable or stressful, there is no reason for them to want to participate (Ioannou-Georgiou, 2011).

c) The assessment tasks and overall assessment system should be child-friendly. Assessment tasks should reflect the cognitive development and particular characteristics of the YL. Moreover, reporting of progress and development should be carried out in ways that are clear and meaningful to YLs (Ioannou-Georgiou and Pavlou, 2003).

d) The assessment system should be criterion-referenced. YLs should see themselves develop according to their own abilities, celebrate their successes and tackle their challenges without having to be compared to others.

e) The assessment system should involve multiple assessments and use a variety of assessment tools. This is essential so as to give learners the chance to show their progress at different points in time, offer opportunities for progress to be discussed and reflected on as well as to see the outcomes of their efforts. This implies that assessment should be formative and happen frequently using a range of tools.

All the above characteristics can be part of a system for formative assessment as well as part of a system which functions external monitoring and accountability. The assessment system, which is shown in the Figure 2 below, proposes portfolio assessment as the main assessment form for YLs.

Portfolio assessment can satisfy all the criteria mentioned above, providing flexibility according to context and particular classes, achieving congruence between classes, learners' needs and assessment. Portfolios also involve multiple assessments and various assessment tasks. Finally, portfolios focus on the individuals and the progress they can make based on their own abilities (criterion-referenced) and support the development goal-setting and self-assessment.

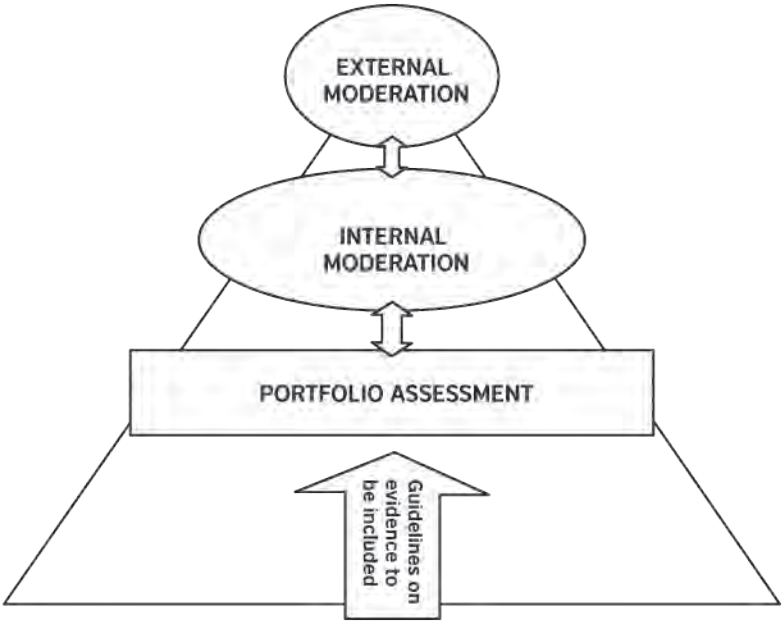


Figure 2 - A model for assessment of YL programmes.

Portfolio assessment can satisfy all the criteria mentioned above. It gives the teachers flexibility to use it according to their context and particular classes, thus allowing them to be consistent between their classes, their learners' needs and assessment. Portfolios also involve multiple assessments and various assessment tasks. Finally, portfolio focuses on the individuals and the progress they can make based on their own abilities (criterion-referenced) while lending itself to the development of self-reflective processes such as goal-setting and self-assessment.

In order to ensure assessment focuses on curriculum objectives and that YLs are achieving the performance indicators specified, portfolios can be accompanied with guidelines and checklists for teachers, while also making it clear that there is plenty of room for teacher and student initiative.

Furthermore, some form of standardisation can be ensured by monitoring which can take place both through peer moderation between teachers and through moderation by Senior Teachers. External moderation of random sample portfolios can take place by school inspectors and also by parents.

A main requirement for the success of the proposed system is teacher training in assessment literacy and training of inspectors and senior teachers in how to moderate the assessment portfolios. It is also important for parents to be informed about the assessment system and trained in how they can help review their children's work.

Finally, in terms of monitoring and assessing the educational programme – and not individual students – external summative exams, locally designed to match the curriculum, could be administered to a random sample of students and/or schools at annual intervals. The exam results would be used for assessing the educational programme and results could be fed back to relevant schools for use in terms of feedback and improvement.

Conclusion

This paper discussed the rise of YL programmes around the world and the need for accountability. It went on to analyse different assessment types as regards their suitability to YL programmes. Finally, the author put forward a range of criteria for assessment of YLs and proposed a model of assessment which can be implemented in YL programmes and which is argued to support the development of a learning atmosphere beneficial to learners, while also satisfying stakeholders' requirements for accountability.

References

Black, P., and William, D. (1998). Inside the black box: Raising standards through classroom assessment. ***Phi Delta Kappan, 80 (2), pp. 139-18.***

Cyprus Ministry of Education and Culture. (2010). ***Αναλυτικά Προγράμματα για τα Δημόςια Σχολεία τησ Κυπριακήσ Δημοκρατίασ*** (National Curricula for the State Schools of the Cyprus Republic).

Clark, A. (1998). ***Gender on the Agenda: Factors Motivating Boys and Girls in MFLs***. London: Centre for Information on Language Teaching and Research

European Commission (2005). A new Framework for Multilingualism. Brussels. European Commission.

European Commission (2006). ***Key competences for lifelong learning***. European Communities

European Commission (2008). ***Multilingualism: an asset for Europe and a shared commitment***. Brussels. European Commission.

Ioannou-Georgiou, S. (2011). ***Voices from CLIL Primary and Preprimary Classrooms***. Paper presented at the 45th IATEFL conference, Brighton.

Ioannou-Georgiou, S., and Pavlou, P. (2003). ***Assessing Young Learners***. Oxford University Press.

Kissau, S. (2006). Gender differences in motivation to learn French. ***Canadian Modern Language Review / La Revue canadienne des langues vivantes, 62 (3), pp. 401–422.***

Loulidi, R. (1990). 'Is Language Learning Really a Female Business?' ***Language Learning Journal, 1, pp. 40-43.***

MacNaughton, G. and Davis, K. (2009). ***Race and Early Childhood Education: An International Approach to Identity, Politics and Pedagogy***. Palgrave Macmillan, New York.

Smith, M. L. (1991). 'Put to the test: The effects of external testing on teachers.' ***Educational Researcher, 20, pp. 8-11.***

5

Reconfiguring assessment to promote productive student learning

David R. Carless

Abstract

This paper focuses on the relationship between testing and student learning. I develop the notion of learning-oriented assessment comprising three elements: appropriate assessment task design; student involvement in assessment through peer- and self-evaluation; and dialogic feedback. I illustrate aspects of these components with examples from schools in Hong Kong. Implications for policy and practice are drawn out and challenges for implementation addressed.

Introduction

The general theme of this paper is to explore how synergies might be developed between summative and formative assessment. In short, formative assessment is to do with eliciting and interpreting evidence, so as to enhance instruction and improve student learning. Formative assessment has been shown to be a highly promising strategy for improving student learning (Black & Wiliam, 1998; Black et al., 2003; Brookhart 2007) yet successful implementation in Confucian-heritage (and other) settings is far from easy in view of the dominance of the summative paradigm (Carless, 2005, 2010; Kennedy et al., 2008). A proposition is that we need contextually-grounded practices which acknowledge the realities of specific socio-cultural settings (Carless, 2011).

A potential way forward is the concept of learning-oriented assessment (Carless, 2007). Learning-oriented assessment is premised on the notion that all assessment should be focused on the development of productive student learning. Three integrated elements of learning-oriented assessment are proposed in Figure 1 below.

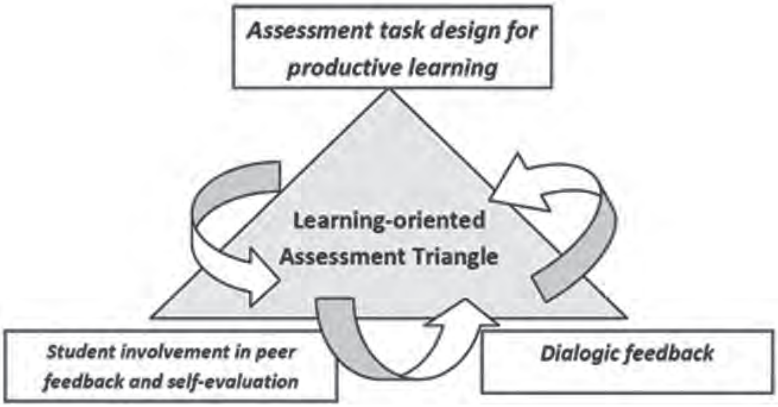

Assessment task design for productive learning
Learning-oriented Assessment Triangle
Student involvement in peer feedback and self-evaluation
Dialogic feedback

Figure 1 - The Learning-oriented Assessment Triangle.

Component 1: Assessment task design for productive student learning
Student behaviours are directly influenced by the assessments they are undertaking. Assessment tasks should be designed to stimulate productive learning practices amongst students e.g. develop long-term learning dispositions rather than short-term memorisation of material.

Component 2: Student involvement in peer feedback and self-evaluation
Assessment should involve students actively in engaging with criteria, quality, their own and/or peers' performance. This process is facilitated by student analysis of exemplars, work of peers, and their own work in progress.

Component 3: Dialogic feedback
Dialogic feedback is defined as interactive exchanges in which interpretations are shared, meanings negotiated and expectations clarified. Through entering into dialogue with peer and teachers, students begin to develop a sense of what quality work entails. Unless students are developing this notion of quality, their ability to make sense of and use the limited and encrypted feedback teachers provide is seriously constrained.

This framework was designed as a tool to cast light on assessment and learning in higher education, although it may also carry implications for other sectors. The remainder of the paper focuses on the school level, drawing on examples and themes in Carless (2011).

Formative use of summative tests

Congruent with learning-oriented assessment is the development of productive relationships between formative and summative assessment. An example is the notion of the 'formative use of summative tests' (Black et al., 2003), hereafter FUST, or to put it more precisely, the formative use of a test designed principally for summative purposes. FUST is focused on how test follow-up can be used to develop ongoing student learning capacities.

FUST recognises the realities of teachers and students lives in that they need to engage actively with both summative and formative assessment. FUST is focused on using information from tests to advance student learning, so has the potential to contribute to a positive relationship between summative and formative assessment. This conception can encourage teachers and students to view test data, not just in terms of grades, but also in relation to rectifying student learning difficulties or supporting students to learn from their test performance. As Brookhart (2001) has shown, successful students take advantage of both summative and formative information to improve their learning.

For FUST to be exploited optimally it requires the right kinds of tests: those that are aligned with curriculum aims; those that are instructionally sensitive (Popham, 2008); and those that promote mastery as well as performance (cf. Dweck, 2000).

Selected findings

Two inter-related projects, reported in Carless (2011), used classroom observations and interviews with teachers and students to probe the implementation of FUST in primary school English as a Foreign Language classes.

Interview data enabled us to probe the interplay between teacher perceptions and their classroom practices. Teachers' potential to develop further the implementation of FUST related to a complex interplay between multiple factors, including: their background, training and experience; beliefs and understandings related to testing and formative assessment; the extent of satisfaction with existing practices; the pedagogic priorities in their own school context; and their lives outside school, which sometimes impacted on their energy and commitment to renew classroom practices.

Interviews with students revealed generally positive orientations towards working with their peers in test preparation and follow-up, except for some cases of conflict between group members. The affective impact of testing on students had powerful impacts on them, and for lower achieving students this could be a major source of discouragement.

Classroom observations revealed a number of test follow-up strategies. The traditional practice of explaining correct answers by 'going through' the test paper appeared to have modest impact on students because the grades had already been awarded and there was little incentive to engage with further teacher input or exhortations.

A strategy employed by several of the teachers was the re-teaching of content. Re-teaching could involve either providing input in similar ways to before as a recapitulation, or it could include a more varied approach which used different teaching strategies. It seemed that the latter was the more promising approach because it often involved teacher reflection on why a teaching strategy had not achieved its full aim, and pragmatically it often included more variety than traditional teacher-fronted whole-class instruction.

Follow-up strategies which carried potential for enhancing student learning and affect were focused on students as active participants in test follow-up. In class, peer co-operation was a key strand of this and was exploited in various ways: students co-operating in pairs or small groups to work out answers or develop correction sheets; and students sharing test preparation and test-taking strategies. Two of our case study schools also developed out of class, across age peer tutoring programs in which older students tutored younger ones with a particular focus on revision and test preparation.

Another potentially useful strategy was engaging students in self-evaluation through written self-reflections on test performance: what they did well, what they were less successful in and how they sought to improve. Whilst there was a limitation that some students wrote relatively trivial comments, such as “Study harder” or “Listen to the teacher”, such processes had potential to show students that a test was not just an end-point but part of ongoing learning.

Underpinning these elements were various contextual and cultural factors pertaining to the Confucian-heritage setting of Hong Kong, including the expected roles of teachers and students; the dominance of examination-oriented education; and collectivist notions encouraging positive relationships with peers.

Contextually-grounded approaches to formative assessment

I propose a possible way forward for learning-oriented assessment which involves the development of contextually-grounded approaches (Carless, 2011). This perspective views the implementation of formative assessment through a socio-cultural lens. It proposes that different variations of formative assessment are needed for different settings. Most of what has been written about formative assessment comes from major Anglophone countries (UK, US, Australia and New Zealand); or from Western Europe. I argue that in Confucian-heritage settings, we need different formative assessment strategies from this literature (Carless, 2011). I call this a contextually-grounded approach in that it is grounded in and derived from the interplay between existing indigenised practices and those found in the Anglophone literature.

A contextually-grounded variation of formative assessment is based on four inter-related principles. First, it has a basis in the existing *beliefs* of teachers and learners, grounded as they are in a particular socio-cultural setting. Second, it uses as a starting-point the existing classroom assessment practices of teachers and seeks ways to build on and enhance these. Third, it takes a realistic and pragmatic view of the extent of formative assessment implementation which is feasible in the selected context under discussion. This is intended to seed starting points for further development of formative assessment practice. Fourth and congruent with the above, it acknowledges the role of high-stakes and other summative assessments in Confucian-heritage settings, cognisant of the major impact these assessments have on students and teachers. From these realities, contextually-grounded formative assessment seeks to find ways in which teachers and students can act together in ways which can advance learning. The emerging starting point can form a basis for continued development towards more extended forms of formative assessment.

It is suggested on the basis of the data from primary schools and the related arguments advanced in chapters 6 and 7 of Carless (2011) that a contextually-grounded version of formative assessment feasible in Confucian-heritage settings might particularly involve two main strands. First, it builds on the strategy of FUST to denote strategies which aim to make test follow-up a rich learning experience for students. Second, it emphasises peer co-operation and peer support through activities, such as peer tutoring which build on the collectivist orientation of Confucian-heritage settings.

Conclusion

Some implications arise. First, a potential strength of FUST is that it can be integrated with other formative assessment strategies, such as peer and self-assessment. Implications for practice are that students need to be supported to work effectively in groups and provided feedback on how to develop their self-evaluation capacities further. Implications for policy are that all assessment (including high-stakes tests) needs to consider carefully its impact on the development of productive student learning dispositions. This carries a further assumption that more communication around assessment is needed, and in particular the further development of assessment literacy of all stakeholders, including ministry officials, policy-makers, teacher educators, parents, teachers and students.

Acknowledgement: This research is supported by two General Research Fund grants from the Research Grants Council of Hong Kong: HKU 7478/06H & HKU 7406/09H.

References

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). ***Assessment for learning: Putting it into practice.*** Maidenhead: Open University Press.

Black, P. and Wiliam, D. (1998). Assessment and classroom learning. ***Assessment in Education, 5(1), pp. 7-74.***

Brookhart, S. (2001). Successful students' formative and summative uses of assessment information. ***Assessment in Education, 8 (2), pp. 153-169.***

Brookhart, S. (2007). Expanding views about formative classroom assessment: a review of the literature. In J. McMillan (Ed.), ***Formative classroom assessment: from theory into practice, pp. 43-62.*** New York: Teachers College Press.

Carless, D. (2005). Prospects for the implementation of assessment for learning, ***Assessment in Education, 12(1), pp. 39-54.***

____ (2007) Learning-oriented assessment: conceptual basis and practical implications. ***Innovations in Education and Teaching International, 44(1), pp. 57-66.***

____ (2010). Classroom assessment in the Hong Kong policy context. In B. McGaw, E. Baker, & P. Peterson (Eds.), ***International Encyclopedia of Education*** (3rd edition). Oxford: Elsevier.

____ (2011). ***From testing to productive student learning: implementing formative assessment in Confucian-heritage settings.*** New York: Routledge.

Dweck, C.S. (2000). ***Self-theories: Their role in motivation, personality, and development.*** Lillington, NC: Taylor & Francis.

Kennedy, K. J., Chan, J. K. S., Fok, P. K., & Yu, W. M. (2008). Forms of assessment and their potential for enhancing learning: Conceptual and cultural issues. ***Educational Research for Policy and Practice, 7(3), pp. 197-207.***

Popham, W.J. (2008). ***Transformative Assessment.*** ASCD: Alexandria.

6

Bridging learning and assessment through readers' theatre

Kaseh Abu Bakar

Abstract

This paper describes an ongoing initiative in using Readers' Theatre (RT) as an assessment activity to reinforce the development of reading and the application of linguistic skills in a foreign language classroom. The author suggests that with a well-developed scoring rubric that takes into account the learning outcomes of the course and the requirements of the performance task, and the implementation of a well-developed washback plan, RT as a performance-based assessment can indeed drive learning and assess learners' achievement.

1. Readers' Theatre as a Learning Activity for L2 Reading

Readers' Theatre (RT) is a form of creative drama in which actors read their lines rather than memorise them. Being a drama, it is a means for linguistic reinforcement in language classrooms (McRae 1985). RT supports L2 reading development as it brings together various forms of reading and their respective advantages. The distinct feature of RT is oral reading, which is important for developing a feeling for the sounds of the target language, encourages students to read in 'meaningful mouths' (Rivers 1981), and heightens their sensitivity to the morphological patterns of the language (Larkin 1995). Rather than the mechanical and meaningless reading as students take turns to read aloud in the classroom, RT creates an authentic oral reading task as learners have to produce a comprehensible performance. And before this can happen, they need to comprehend the text by engaging in intensive reading, applying reading strategies and attending to linguistic features as they construct meaning. In preparing for a smooth presentation, RT inevitably provides a legitimate reason for repeated reading, a practice that is proven to lead to significant gains in reading fluency (Rasinski, as cited in Prescott 2003). The collaborative nature of RT as a group assignment lessens dependence on the teacher, helps learners overcome reading anxiety and pushes them to construct meaning collectively, verbalising their comprehension processes as they go along, which is also an advantage to the instructor in checking their comprehension and identifying problems. RT also caters to learner-reader emotions. Being a learner-centred activity, it is engaging and appealing. It allows individual differences in learning and expression, it quietens their stage fright as they can hide behind the papers, and all these in turn boost their interest and confidence in reading.

2. Developing RT as an Assessment Tool

The fact that RT is a learner-centred performance activity renders it suitable for a performance-based assessment. In designing RT as a performance-based assessment that drives learning, several validity concerns need to be addressed. Apart from the appropriateness and validity of the performance task in relation to the intended outcome, another critical concern is rubric validity, which is essentially, "...the appropriateness and validity of the criteria and descriptors for discrimination in relation to that task" (Wiggins 1998, 164). A valid rubric should specify a clear set of assessment criteria linked to the course learning outcomes and the most important dimensions of performance, and provide authentic and effective descriptors for discriminating between levels of performance (Wiggins 1998). This subsection describes the development of the scoring rubric for RT in a remedial Arabic as a foreign language classroom, which I believe could also be applied with adaptations to ESL and other foreign language contexts.

The Remedial Arabic 1 course is designed for post-graduate students at the Faculty of Islamic Studies, Universiti Kebangsaan Malaysia, who are non-native Arabic speakers and who do not meet a specified standard in an Arabic placement test. The main objective of the course is to guide learners in interacting with contemporary Arabic texts effectively and confidently. In addition, basic morphological and grammatical skills which are considered essential for supporting the development of reading proficiency are also emphasised. The texts which are used in the class are contemporary texts representing various disciplines in Islamic Studies as well as a few texts for pleasurable reading, such as jokes and short stories.

Multiple indicators of achievement are used to assess the students. RT constitutes 15% of the total and directly assesses the students' effective performance in the oral reading of a contemporary short story. The learning outcomes to be assessed are competence and confidence in interacting with and comprehending contemporary Arabic texts and competency in applying essential morphological and syntactical skills for successful comprehension. The RT task requires i) clear, fluent pronunciation, intonation and other paralinguistic features; ii) accuracy in assigning diacritic sounds to nouns, verbs, and other word classes; and iii) appropriate proxemics and creativity in performance. The blend of the two sets of requirements results in three assessment criteria for the RT performance: fluency of pronunciation and suitability of intonation; accuracy in morphology and grammar, and creativity and suitability of performance to the context of the story (Figure 1).

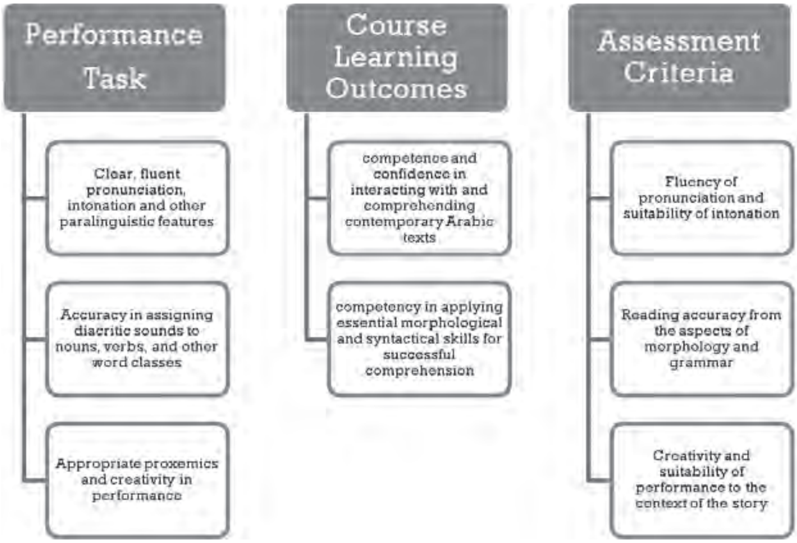

Performance Task
Clear, fluent pronunciation, intonation and other paralinguistic features
Accuracy in assigning diacritic sounds to nouns, verbs, and other word classes
Appropriate proxemics and creativity in performance
Course Learning Outcomes
competence and confidence in interacting with and comprehending contemporary Arabic texts
competency in applying essential morphological and syntactical skills for successful comprehension
Assessment Criteria
Fluency of pronunciation and suitability of intonation
Reading accuracy from the aspects of morphology and grammar
Creativity and suitability of performance to the context of the story

Figure 1 - Identifying the important dimensions of performance, target criteria of the learning outcomes and the assessment criteria.

Indirectly, these criteria allow us to infer from the scores that students have acquired certain fluency in their pronunciation and intonation of Arabic sounds; that they can accurately apply morphological and syntactical rules to the script; and that students understand the story well enough to provide a creative performance. These requirements and criteria represent the learning processes and outcomes designed to be part of the students' learning experience. Three levels of performance were further described for each criterion (Table 1).

Criteria	Scale	Description
Fluency of pronunciation and suitability of intonation.	8-10	Pronunciation and intonation are close to those of native Arab speaker. Very clear and easy to follow by audience.
	4-7	Pronunciation and intonation are less close to those of a native Arab speaker, but comprehensible enough for a sympathetic listener. Sometimes not clear, causing slight difficulty for audience to understand the presentation.
	1-3	Pronunciation and intonation not at all close to those of native Arab speaker. Not clear, causing difficulty for audience to follow presentation.
Reading accuracy from the aspects of morphology and grammar.	8-10	Accurate reading from aspect of morphology and grammar. Very helpful for listeners to understand the story.
	4-7	Reading sometimes does not follow rules of morphology and grammar. Listeners sometimes face difficulty to understand the story.

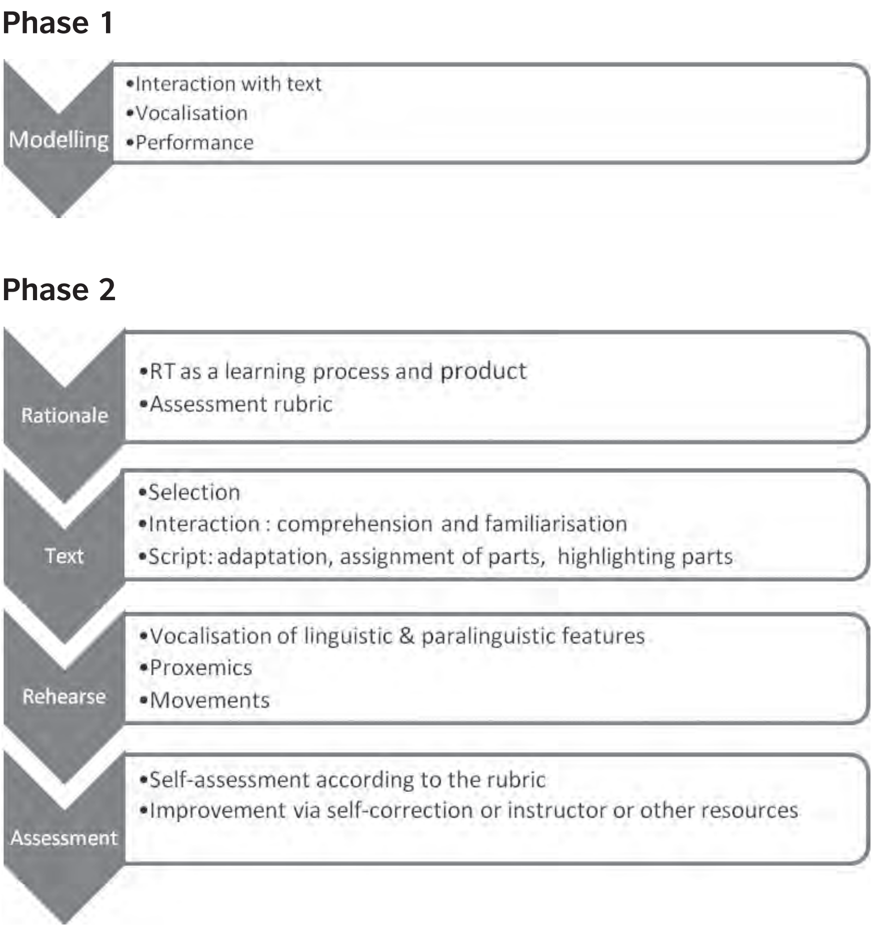
Creativity and suitability of performance to the context of the story.	1-3	Frequently breaches rules of morphology and grammar. Very difficult for listeners to understand the story.
	8-10	Smooth presentation. Body movements and facial mimics very appropriate and helpful for understanding.
	4-7	Presentation was sometimes discontinuous. Body movements and facial mimics were not much used, and if used, were not so appropriate or helpful for understanding.
	1-3	Presentation was frequently discontinuous. Body movements and facial mimics were not used or not appropriate and not helpful for understanding.
Total Marks	**30**	**Assessment percentage 15%**

Table 1 - Assessment Rubric (reproduced with improvements from Kaseh et al 2011).

3. Enhancing the Positive Washback of RT

The purpose of introducing RT in this course was to promote learning practices that otherwise would not take place, a notion referred to as washback (Messick, Bailey 1996). Figure 2 summarises the model that was employed to ensure the positive effect of RT as a learning-cum-assessment tool.

In the modeling phase, it is important that the instructor demonstrate an effective interaction with a short story, help learners identify the context, settings, characters, problem, events, climax, solution/ending, vocalise a good oral reading, and whenever possible, demonstrate an RT performance of the story. The next phase shifts the focus on learners. The teacher needs to communicate the rationale of the activity and go through the assessment criteria with the learners. It is essential that learners become aware of the importance of the process and product and the learning values of RT. Next, is the text phase. Assign a pre-selected short story to each group or help it to select a suitable one with sufficient characters and events. In our emphasis on assigning diacritic marks as a demonstration of learners' application of Arabic morphological skills, we request that learners select a text that does not come with diacritic marks. At this stage, learners familiarise themselves with the story and engage in constructing meaning. These are done through group discussion and facilitated by the instructor. Instructor guidance is often needed to process the cultural nuances and some linguistic features of the text. Learners then move on to adapting the text for RT, assigning roles and parts to group members, and highlighting individual parts.



Phase 1

Modelling
•Interaction with text
•Vocalisation
•Performance

Phase 2

Rationale
•RT as a learning process and product
•Assessment rubric
Text
•Selection
•Interaction : comprehension and familiarisation
•Script: adaptation, assignment of parts, highlighting parts
Rehearse
•Vocalisation of linguistic & paralinguistic features
•Proxemics
•Movements
Assessment
•Self-assessment according to the rubric
•Improvement via self-correction or instructor or other resources

Figure 2 - The approach used in ensuring positive washback of RT.

The rehearsing phase is when learning occurs most, and is therefore the most crucial. This is the stage where learners get the opportunity to learn and practise their vocalisation in particular, together with proxemics and movements. It is useful to allocate some 10 to 15 minutes of each lecture prior to the actual presentation for learners to practise their oral reading. Learners often find it helpful to imitate native Arab pronunciation from U-tube or Arabic television channels. At this stage, the instructor can further help learners to discover and correct their mistakes. Throughout the practice sessions, learners self-assess and improve their performance and progress against the assessment rubric.

4. Other Assessment Issues

Face Validity

Any form of assessment, be it summative or formative, needs continuous evaluation to ensure that the assessment and the intended outcomes are commensurate (Genesee & Upshur 1996). The learners surveyed in the course agreed that RT was an appropriate assessment tool for assessing the learning outcomes of the course, indicating high face validity. In addition, they strongly agreed that RT has improved their pronunciation, fluency of Arabic oral reading, application of morphological and syntactical knowledge, as well as interest and confidence in reading Arabic texts (Kaseh et al 2011).

Would it matter if everyone scores equally well?

Ideally an assessment should discriminate among individuals of different levels of performance. While differences are often observed in all three criteria, especially with regards to fluency and accuracy, a hypothetical question would be: what if everyone scores equally high in one or more of the criteria? Would this indicate that the descriptors are not discriminating enough, and therefore the rubric lacks validity? I would say that if the equally high scores are the result of students practicing their RT according to the rubric, and ensured that they have met the highest targeted achievement, then there really is no reason for alarm as it indicates that positive washback has occurred as desired. However, if the similar high scores result from referee lack of ability to discriminate, then it becomes a valid concern to be addressed.

Enhancing the Rubric Validity

The current rubric was developed on a hypothetical basis and has not been subjected to expert review. At least three tasks need to be done in order to enhance the rubric validity: firstly, the criteria need to be revisited; secondly, the descriptors need to be revisited, based on actual samples of different quality performances (Wiggins 1998), and thirdly, the rubric should be subjected to expert review.

5. Conclusion

RT is a performance-based assessment that can be used to assess and reinforce the learning experience of linguistic and skill-based learning outcomes in foreign language classrooms. The effectiveness of RT as an assessment-cum-learning activity depends on i) developing a rubric that blends task and target criteria that measures the learning outcomes, and using the rubric as the basis for rehearsals and actual performance; and ii) planning and executing the processes that would realise the intended positive washback.

References

Bailey, K. (1996). Working for washback: a review of the washback concept in language testing. ***Language Testing, vol. 13 no. 3, pp. 257-279.***

Kaseh Abu Bakar, Hakim Zainal, Maheram Ahmad & Md. Nor Abdullah. (2011). The Empowerment of Postgraduate Students in Arabic: The Case of Faculty of Islamic Studies, Universiti Kebangsaan Malaysia. ***Procedia Social and Behavioral Sciences (18), pp. 481-490.*** www.elsevier/locate/procedia

Larkin, M. (1995). The role for close reading in the elementary Arabic curriculum. In M. Al-Batal (Ed.) ***The teaching of Arabic as a foreign language: Issues and directions, pp. 157-173.*** Provo: American Association of Teachers of Arabic.

Genesee, F. & Upshur, J.A. (1996). ***Classroom-based evaluation in second language education.*** Cambridge: Cambridge University Press.

McRae, J. (1985). Using drama in the classroom. Oxford: Pergamon Institute of English.

Messick, S. J. (1996). Validity and washback in language testing. ***Language Testing, vol. 13 no. 3, pp. 241-256.***

Prescott, J.O. (2003). The power of Reader’s Theatre: an easy way to make dramatic changes in kids’ fluency, writing, listening and social skills. ***Scholastic Instructor, January/February 2003.*** Retrieved 27/06/2011 http://teacher.scholastic.com/products/instructor/readerstheater.htm

Wiggins, S. (1998). ***Educative assessment: Designing assessments to inform and improve student performance.*** San Francisco, CA: Jossey Bass.

7

Worrying for exams or learning to learn lives: An auto/ethnographic approach of assessment through journaling

Kashiraj Pandey

Abstract

As teachers, when we talk about evaluation or assessment in a writing classroom, we need to go beyond simply ranking, assigning grades or completing report cards, to consider something more complex, arbitrary, subjective and contextual. Marks are important as they are part of the route to university and the job market. But exploring students' reflective journals to understand hidden attributes is also useful to understand what students do not know and where to help them. Realising that it is equally beneficial for teachers to review their accomplishments through students' reflective journals, we can also improve, redesign, or revise the style of teaching with the students in mind. Hence, this paper examines the relationship between teaching and testing from alternative integrative perspectives to the reflective mode of learning in creative writing. I have tried to clarify a considerable amount of evidence to justify that journaling can improve various aspects of students learning.

Setting the Scene

As teachers, we want our students to demonstrate what they really know. But, how often do we think about what questions we ask and whether our questions have stimulated the learners to produce something of their own? For me, setting questions is a quest to make room for creative and innovative answers, while exams work as powerful instruments to demonstrate learning, for students and myself. I know that some students who have good memories can get good results in dictation but cannot use the word/s in sentences. Many students memorise vocabulary items before a test but they are soon forgotten. The rubric, "a description of specific level of performance within a performance scale (Gullickson 2003, p 231)" represents what I want the learners to learn, where testing is associated with intrinsic and extrinsic values that prepare the learner to work with new ideas. I prefer to give less value to the achievement of high scores when compared to the development of deep understanding of the material/s provided. When our students fail to learn, no matter how strictly we conduct the exam or prepare them in the classroom, our focus should lie more in how to motivate them and improve learning.

Journaling - What is it?

Journaling in its various forms is a means for recording personal thoughts, daily experiences, and evolving insights (Hiemstra 2001). This reflective process often evokes conversations with self and a real or even an imagined other person making the practitioners able to review or reread the earlier reflections with a progressive clarification of possible insights. Writing provides the students with abundant practice of examples of the subtle and complex uses of grammar and vocabulary of a given language. When I have practised journaling with my students, as a tool in language learning in an undergraduate classroom, I have found great value in the process to improve language skills. But even within this approach which helps to stimulate creativity and understanding, testing as the traditional means of evaluation is still a matter of concern.

On your marks… Get Set… Go!

Starting with familiar settings, auto as the self "I" and ethno as the "culture/s around us", my focus lies in whether we expect to see transformation in only the students, or in both the students and the teacher. Narratives based on reflection/s may not always have concrete evidence, and auto-ethnography, as Creswell (2002) claims is, "a reflective self-examination by an individual set within his or her cultural context" (p 438). Ellis and Bochner (2000) think that the "auto-ethnographical genre of writing and research … displays multiple layers of consciousness, connecting the personal to the cultural," (p 739). I am using autoethnography as the methodology, method and genre of the "self" inquiry in my work/s. By nature, being a university teacher, I too always happen to read, write, and interact with students, preparing them to produce an individual portfolio or a journal every semester. Proposed as an 'insider's' methodology (Luitel 2009), I use autoethnography in such a way that my personal and professional experiences become the key basis of an inquiry. Reflective journaling, as a tool towards transformation, helps us to, "reveal what … writers have learned, examine how writers have learned to express themselves in journals, or find out how journals can help other people to learn," (Boud 2001, p 10).

Benefits

Talking about the reward of reflective writing, "teachers have found that practising diary writing with students may contribute to the learning process as the students are encouraged to continue reflecting on their learning experiences and to try discovering that they might otherwise may not see," (Van Manen, 1990, p. 73). In our case, it has potential positive outcomes related to the promotion of deep learning with increased awareness and improved thoughtfulness before, during, and after the practice.

The principle challenge is how do we know who is being/has been transformed. Unlike more traditional methods, focused on exams more than 'learning to learn about lives', by interacting with my students and analysing the related outcomes (comparing the writings before and at the of the end semester), I have observed differences and 'transformation'. We (student and teacher) are different from what we were before. Journaling has worked as an enabling factor for me and my students to enhance wider imaginative possibilities. I have come to the realisation that the principal benefits of the approach are creativity, personal growth and development in learning with improved self–confidence.

Teaching (Educating) and Testing

As an educator, I engage in critical reflection when it comes to my own teaching and evaluation practices asking myself such questions as why do I teach or train the way I do; what are my goals for the learners and myself as a professional educator; is critical reflection something that needs to be fostered in the context in which I teach or train; and do we have or need adequate orientation to testing.

For me, the best teaching is critically reflective with constant scrutiny of assumptions about teaching and testing. “Many teachers are so focused on teaching that they do not have time to notice if their students are learning,” (Jesus & Bastidas 1996). I frequently revisit critical questions in my professional practice: therefore, am I really interested in transforming the self, the students and society? Do I only teach or somehow also create a learning environment, and how do the students experience their learning while, “the teacher benefits as well, for the permanent record of writing gives a rich, ongoing picture of students’ development as individuals, thinkers, and writers” (Peyton & Reed 2003, p 107). My testing leads toward preparing to know what the students do not know and to find out where to assist to prepare them for the job market, at least, and let the world do the rest. Therefore, reading their reflections makes me realise where I still could have done/do better. Students too revisit their own status in the same ways. I take it as a great opportunity, a journey to explore ourselves, both as teacher and student. My mental process as a test taker is to consider whether the learning (somehow) resembles the situation that the student has to deal with in real life, and makes the connection between what they know and what they want to say.

Conclusion

Journals as the product of subjective, open-ended, and inquiry based reflection, present a unique challenge for evaluators, as they are linked with a set of standards and control. Journals automatically monitor and improve progress, allowing students to self-regulate their own learning towards the desired goals, and as teachers and evaluators, we can assess students’ understanding to decide what support and practice they need more of as revealed in their reflective journals. This can be done through a range of regular classroom tests and quizzes, checklists, writing narratives, project works and/or case studies. This way, testing needs to link to personalised learning, performance accountability, and workplace requirements to explore how students use the knowledge they have, as today’s students are the successful workforce of the future and reflective practice is evidence of life-long learning. Therefore, believing rote memorisation and the ability to get “correct” answers as hindering factors in the exams, I would propose that evaluating our own and our students’ progress through journaling encourages creativity and reflection (with an automatically conscious flow of mind, just as a good driver does not have to consciously think about the process of changing gears). We need to think about how to train a whole person rather than limiting it to an arbitrary timeframe of three or four hours of *closed door* testing.

References

Boud, D. (2001). “Using Journal Writing to Enhance Reflective Practice”. ***New directions for adult and continuing education, vol. 90, no. 10, pp. 9-18.***

Bochner, A. P. (2000). “Autoethnography, personal narrative, reflexivity”. In N. K. Denzin & Y. S. Lincoln (Eds.) ***Handbook of qualitative research***, Sage, Thousand Oaks, CA.

Creswell, J. W. (2002), ***Research design***, 2nd edn, Sage, California.

Ellis, C. & Bochner, A. (2000). “Autoethnography, personal narrative, reflexivity: Researcher as subject.” In N. Denzin & Y. Lincoln (Eds.), ***The handbook of qualitative research (2nd ed.), pp. 733-768***. Newbury Park, CA: Sage.

Gullickson, A. et. al. (2003). ***The student evaluation standards***, Corwin Press and ETS, California.

Field, J, (2011). “The elusive skill: how can we test second language listening validity?” In Powell-Davies, P. (Ed.). ***New Directions in Assessment and Evaluation Symposium.*** British Council Malaysia.

Hiemstra, R. (2001). “Uses and benefits of journal writing”. In L.M. English and M.A. Gillen (Eds.) ***Promoting journal writing in adult education (New Directions for Adult and Continuing Education), pp. 19-26.*** Jossey-Bass, San Francisco.

Jesus, A. & Bastidas, A. (1996). “Teaching portfolios as assessment tools”. ***Forum, vol. 34, no. 3 & 4, pp. 24-28.***

Law, B. & Eckes, M. (2010). ***Assessment and ESL***, 2nd edn. Portage & Main Press, Chicago.

Luitel, B.C. (2009). ***Culture, worldview and transformative philosophy of Mathematics education in Nepal: A cultural philosophical inquiry.*** (Unpublished doctoral dissertation), Curtin University of Technology, Australia.

Mezirow, J. (1997). ***New directions for adult and continuing education 97, Issue 74, pp. 5-12***. Available from: http://scholar.google.com/scholar?q=Transformative+Learning:++Theory+to+Practice+Jack+Mezirow&hl=en&as_sdt=0&as_vis=1&oi=scholart. [15 January 2010]

Peyton, J. & Reed, L. (2003). ***Dialogue journal writing with non-native English speakers***: a handbook for teachers. Alexandria, VA: Teachers of English to Speakers of Other Languages, Inc.

Schon, D. A. (1983). ***The reflective practitioner***. Basic Books, New York.
---- (1987) ***Educating the reflection practitioner: Towards a new design for teaching and learning in the profession***, Jossey-Bass Publishers, San Francisco, CA.

Van Manen, M. (1990). ***Researching lived experience: human science for an action sensitive pedagogy***. Suny Series in the Philosophy of Education. SUNY Press, NY.

Whitehead, J. (2010). “As an Educator and Educational Researcher, How Do I Improve What I Am Doing and Contribute to Educational Theories That Carry Hope for the Future of Humanity?” ***Inquiry in Education: Vol. 1: Issue 2, Article 2***. Available from: http://www.actionresearch.net/writings/writing.shtml>. [1 January 2010].

8

The development of a student learning outcomes–based accreditation model in institutional and programme accreditation in Taiwan Higher Education

Angela Yung-Chi Hou

Abstract

Student learning is a central concern in higher education and accreditation nowadays. Many institutions, programmes, and accrediting organisations are hearing a similar request about student learning outcomes to provide concrete evidence of student academic achievement in higher education and to report on this evidence in a manner that is readily understandable to the public at large. Hence, the public, higher education community, policy makers, and students increasingly seek to use such information as an integral part of making judgments about the quality of accredited institutions and programmes. The main purpose of the paper is to examine recent Taiwan educational policy trends that emphasise learning outcomes and quality assurance.

Introduction

Today the rapid expansion of higher education institutions throughout the world and their increasingly market-based orientation has led students, parents, higher education, employers and governments to have more interest in the quality of universities and colleges. Universities and colleges are beginning to take on accountability towards schools and societies as enterprises do, and are increasingly presenting institutional effectiveness to the general public. Hence, quality assurance mechanisms and international benchmarking, which emphasise output monitoring and measurements and systems of accountability and auditing, have become more popular worldwide (Marginson, 2007).

How to establish and maintain quality in the more than 160 higher education institutions in Taiwan has become a major concern for all stakeholders. In order to improve quality in Taiwan higher education consistently, the Taiwan government started to develop a quality assurance system in the 1980s resulting in a decentralised system of quality assurance when the Higher Education Evaluation & Accreditation (HEEACT) body was founded in 2005. Nowadays, all universities and colleges are obliged to be assessed by one of the external quality assurance agencies according to the University Law Revised of 2005. In the first cycle of programme accreditation from 2006 to 2010, HEEACT mainly used input and process indicators, such as faculty quality, financial resources, research and professional performance, to assess the quality of higher education institutions.

Today, many institutions, programmes, and accrediting organisations in Taiwan are hearing requests about learning outcomes from a number of sources to provide concrete evidence of student academic achievement in higher education and to report on this evidence in a manner that is readily understandable to the public at large. In order to respond to legitimate public demand, HEEACT started working toward greater emphasis on student learning outcomes in 2011/2012. Many institutions, policy makers and other stakeholders were invited to discuss with HEEACT how evidence of the attainment of learning objectives can be successfully obtained. Therefore, the main purpose of this paper is to examine recent Taiwan educational policy trends that emphasise learning outcomes and quality assurance.

Development of Learning Outcomes-Based Accreditation

Over the past decade, increasing pressure to demonstrate accountability of higher education had led to the rise of learning outcomes based assessment in many countries. Hence, a debate over how to gather reliable evidence of the student achievement of these outcomes has been growing. According to Wolff (2009, p.84), the focus had made accreditors shift accreditation standards away from, “the use of key input and resources indicators to gain evidence of effectiveness, especially in relation to student learning outcomes.“

The U.S.A. was one of the first nations to focus on learning outcomes assessment. In the mid-80s, U.S. higher education began a so-called ‘assessment movement’, as Ewell stated, “aimed at gathering systematic evidence on student learning outcomes and a call to provide information that enabled institutions to establish clear statement of intended learning outcomes and make the result public,” (Ewell, 2008, p. 42). In the early 90s, over 90% of institutions had an assessment programme under way. At the same time, U.S. regional accreditors played a very prominent role in outcomes assessment. Programme and career-related accreditors also paid increasing attention to evidence of student academic achievement by requesting programmes to develop assessment systems (Ewell, 2008).

In the early 90s, the UK government began to express its concerns about institutional standards. A discipline-based panel was convened by the Quality Assurance Agency to create “subject benchmark statements”, which describe what can be expected of a graduate in terms of abilities, skills, understanding and competence in the subject. In fact, setting up standards and gathering evidence remained a big challenge for UK institutions. In order to assess learning outcomes concretely, other nations, for example, Australia, Hong Kong, developed National Qualification Frameworks to assist accreditors assess if students had achieved the intended learning goals (Woodhouse, 2010).

Accountability mainly aims at improving the fiscal efficiency of an educational organisation. Assessment, on the other hand, is used to focus to a greater extent on the quality of education. Therefore, student-learning outcomes are assumed to be better indicators of institutional quality or effectiveness based on the newly developed concept of "assessment for accountability" in higher education. Hence, notions of quality in accreditation, defined in terms of input and process standards, has evolved into notions of quality based on institutional mission fulfillment over decades, and they are now moving toward student learning outcomes based assessment (Ewell, 2008).

Assessment for Student Learning Outcomes

Student-learning outcomes generally refer to aggregate statistics on groups of students, such as graduation rates, retention rates, transfer rates, and employment rates for an entering class or a graduating class. They represent institutional performance, rather than what and how students learn. With a boarder definition, student learning outcomes now encompass, "a wider range of student attributes and abilities, both cognitive and affective, which are a measure of how their college experiences have supported their development as individuals", which include acquisition of specific knowledge and skills, values, goals, attitudes, self-concepts, world views, and behaviours (Frye, 2009). To summarise, student learning outcomes can be broadly defined as, "something which happened to an individual student as a result of his or her attendance at a higher education institution and/ or participation in a particular course of study" (Ewell, p.5).

When it comes to student learning assessment, several issues pertaining to content, methodology and evidence are raised. Ewell (2001) proposed 4 dimensions of student learning outcomes assessment that an accrediting agency will adopt in terms of 3 dimensions of choices: prescription of outcomes, unit of analysis and focus of review. An accrediting agency should specify the particular learning outcomes for the accredited programmes and institutions and examine the direct evidence of student achievement to assure the quality of learning outcomes. The characteristics of the four dimensions: programme assessment, academic audit, auditing academic standards, and third party certification, are summarised in Table 1.

Based on the 4 dimensions in Table 1, there are a number of problems: such as what kind of evidence should be considered acceptable by an accrediting agency and how will it be collected by the programme and institution. Evidence should not only be relevant but also verifiable by third party inspection, particularly accrediting agencies. Several types of evidence are usually collected from faculty-designed examinations and assignments, performance on licensing or external examinations, portfolios of student work, student satisfaction surveys, interviews, etc. However, there is no guarantee that they represent directly what students learn in universities and colleges.

	PROGRAMME ASSESSMENT	ACADEMIC AUDIT	AUDITING ACADEMIC STANDARDS	THIRD PARTY CERTIFICATION
PRESCRIPTION OF OUTCOMES	Programme and institution	Programme and institution	Student	Student
UNIT OF ANALYSIS	Programme and institution effectiveness	Programme and institution effectiveness	Individual attainment	Individual attainment
FOCUS OF REVIEW	Indirect evidence / portfolios, examination and survey over students.	No direct evidence / learning outcomes are decided by the institution and programme	Direct evidence/ student work products, student career development.	Direct evidence/ licence and certificates

Table 1 - Characteristics of Four Models of Assessment for Student Learning Outcomes. Source: CHEA (2001). Accreditation and student learning outcomes: A proposed point of departure. Washington, D.C. CHEA.

Quality Assurance in Taiwan Higher Education

As higher education enrolment has expanded over the past 20 years to the present 1.3 million students, the public's desire to maintain and increase both "quantity" and "quality" has placed tremendous pressure on the government. In the 1990s, the government began implementing a wide range of comprehensive institutional evaluations with the goal of establishing a non-governmental professional evaluation agency whose purpose was to conduct evaluations of higher education institutions. Until now, three independent evaluation agencies officially chartered by the Ministry of Education have begun to assess three different types of Taiwan higher education institutions: four-year comprehensive colleges and universities, universities of science and technology and technical colleges. A comparison between the three agencies is contained in Table 2 below.

	HEEACT	TWAEA	NYUST
Background			
Starting year	2006	2004	2002
Type	Non-profit Foundation	Non-profit Foundation	Higher education institution
Governance	15 Board members	15 Board members	Research centre (6 staff)
Funding	Ministry of Education	Ministry of Education	Ministry of Education
Content of Quality Assurance			
Nature	Mandatory	Mandatory	Mandatory
Unit	Programme	Institutional/ programme	Institutional / Programme
Scope	76 4-year comprehensive colleges and universities	38 Universities of Science and Technology	40 Technical Colleges (including 2 and 5 year junior colleges)
Process	Self-evaluation / peer review	Self-evaluation / peer review	Self-evaluation / peer review
Standards	5 criteria	5 items in institutional evaluation and 8 items in programme evaluation	5 items in institutional evaluation and 8 items in programme evaluation
Review cycle	5 years	4 years	4 years
Outcome	1. Accredited 2. Accredited conditionally 3. Denial	Rank 1: above 80 points Rank 2: 70~80 Rank 3: 60~70 Rank 4: below 60	Rank 1: above 80 points Rank 2: 70~80 Rank 3: 60~70 Rank 4: below 60
Implications	Governmental Funding / enrolment approved	Governmental Funding / enrolment approved	Governmental Funding / enrolment approved

Table 2 - Comparison among three quality assurance agencies by background and accreditation status. Source: Author.

The example of HEEACT is indicative. Over 800 reviewers from universities and industries are recommended by 47 Programme Planning Committees formed by the Board to conduct evaluations (HEEACT, 2009b). The accreditation standards developed by the HEEACT are as follows: 1. goals, features, and self-enhancement mechanisms, 2. curriculum design and teaching, 3. learning and student affairs, 4. research and professional performance, 5. performance of graduates. There are three review accreditation outcomes - "accredited', "conditionally accredited" and "denial". Those with a status of "conditionally accredited" or "denial" are supposed to be reviewed again one year later to check if all major problems mentioned in the final accreditation report have been solved during the year. A 2-4 page formative and summative report is provided after the assessment with comments based on site visits and the proposed accreditation outcome. In the former one, the strengths and weaknesses of the programme are evaluated against the 5 standards. As to accreditation outcome, a checklist of 5 criteria accompanied with 21 indicators is provided for reviewers. The review panel assesses the performance of the programme on a 6-point scale (6-excellent; 1-not good) based on the items of the checklist. On the bottom of the checklist, the whole panel will suggest the final accredited status based on the evidence provided, send it to the *Preliminary Accreditation Review Subcommittee,* and then the *Accreditation Review Committee* which finalises the accredited status. The list of 5 criteria and 21 indicators is as follows (HEEACT, 2008b):

Item one: Mission, Goals, features, governance, self–improvement system

1. Faculty and students fully understanding mission and goal of the programme
2. Mission and goal related to the institutional development
3. Operation of self-evaluation mechanism
4. Self-improvement system
5. Operation of various steering committees of the programme.

Item two: Curriculum design and instruction

1. Curriculum development and planning
2. Curriculums meeting programme goal and mission
3. Quality and quantity of full-time and part-time faculty and staff satisfying student demands
4. Instruction content following course schedule
5. Faculty development and teaching quality are the centrality of the programme
6. What faculty members teach corresponds to the academic scholarship.

Item three: Student learning and student Affairs

1. Teaching resource satisfying student demands
2. Sufficient faculty resources helping students write theses and dissertations.
3. Student support and services in learning and counseling
4. Tutorial time scheduled
5. Students' opinions being respected and accepted
6. Active interaction between advisors and graduate students.

Item four: Research output and professional performance

1. Research output and professional performance of faculty
2. Grants and research projects received by faculty
3. Faculty members providing professional services for the community.

Item five: Alumni performance

1. Effective channels to contact with graduates and their employment and career tracking.

The final reports are provided for three major stakeholders, including the institutions for self-improvement, the government for resource allocation and students for school selection (HEEACT, 2009).

To date, four rounds of accreditations have been conducted, and the results of the first three and a half rounds have been released. According to the review outcomes of the past four years, the average rate for 'accredited' status is 83.21%, for 'conditionally accredited' 14.03%, and for 'denied' 2.69 %. It is evident that these figures demonstrate that Taiwan institutions are becoming more acquainted with the HEEACT accreditation model, and that they are aiming at self-enhancement, and learning ways to prepare faculty for participation.

The 2011 Institutional Accreditation and 2012 Programmatic Accreditation – Student Learning Outcome based Model

Prior to the establishment of HEEACT, higher education evaluations in Taiwan mainly focused on input and process measures. When HEEACT conducted a programme accreditation exercise in 2006, learning outcomes started to gain increasing attention. In the 2007 Evaluation on Colleges of Science & Technology and Technical Colleges, the item of "student achievement and development" was adopted as one of the standards of programme effectiveness (TWEAA, 2007). Recently, the Ministry of Education announced a new policy that would put greater emphasis on evidence of educational effectiveness and student learning in the upcoming institution and the new cycle of programme accreditations. In order to ensure the levels of the professional knowledge and skills students require in the job market, the Guidelines of "Promoting Student Quality in Postsecondary Education Program" initiated by the MOE in 2009 indicated clearly that all programmes and institutions are required to set a series of core competencies or to hold exit tests for all students (MOE, 2009a). It also stated that all applicants will be reviewed and selected according to a common set of criteria, including teaching quality, student learning effectiveness and curriculum and programme planning. In the dimension of student learning outcomes, applicants are required to provide some evidence pertaining to institutional effectiveness, including freshmen counseling support, core competence development, mechanism for alumni tracks, healthy functions of career planning office, citizenship cultivation, etc. (MOE, 2009b).

Starting in 2011, HEEACT is planning to conduct a comprehensive assessment of over 81 4-year national and private universities and will also continue the second cycle programme accreditation in the following year. There are 5 Items set for the 2011 institutional accreditation:

1. *Institutional mission - what was the institutional long-term development plan?*
2. *Institutional governance and management--how was the organisational management system established?*
3. *Teaching and learning resources--how were educational resources allocated to enhance teaching quality and assure learning outcomes?*
4. *Accountability and social responsibility—what do learning outcomes consist of? How are they achieved?*
5. *Self-enhancement mechanism and quality assurance culture—how do institutions self improve continuously.*

In order to help institutions understand the core of student learning outcomes-based accreditation, there are a number of questions for institutional self review and evidence preparation:

1. *What are students' core competences and professional skills and how are they determined?*
2. *How do institutions and programmes assess students' core competences and professional skills?*
3. *How are curriculum and programmes designed to achieve learning goals?*
4. *How can student learning outcomes be improved if they are not as expected?*
5. *How are student support and services provided?*

In fact, all institutions are encouraged to determine the content of students' core competences and professional skills according to their own mission and characteristics. Most important of all, institutions are supposed to develop a continuous self-improvement mechanism on campus to assure learning outcomes and quality. They are also advised to make a practical and sustainable policy to achieve goals, rather than simply follow standards, in other words, these goals should be achievable with its educational resources (HEEACT, 2010).

HEEACT has emphasised that the differentiation and diversity of universities is fully respected in the first phase of the implementation of the new outcomes-based model. Most importantly, HEEACT will not make a comparison between universities according to the evaluation outcomes. However, universities are encouraged to select measurable learning outcomes, to develop a variety of assessment tools at the course, programme and institutional levels, and to prove that the learning outcomes are met.

It is evident that the goal of the new 'OBA' model applied in the 2011 institutional accreditation is, "to ascertain whether each institution is operating well according to its mission and goals, and to assist the institution to identify itself, to find its strengths and weaknesses, to develop its features, and to engage in self-improvement through each institution's self-evaluation and onsite visits" (HEEACT, 2010, p. 4).

Several universities have taken actions in the development of student learning outcomes, such as establishing clear statements describing outcomes; collecting and interpreting evidence of student performance; routinely modifying the standards, policies, curricular structure and learning support systems based on the opinions of graduates, employers, and student e-portfolio, etc. Take Soochow University, for example. It successfully designed student attributes and competencies in three domains: general education, social and interpersonal skills and professional knowledge. Then the intended generic and professional competency indicators were embedded into curriculum design, stressing the connectivity of the theory and practice. Finally, three domains of student learning outcomes and competency indicators are built into the customised student e-portfolio system, which helps students realise the quality of their learning outcomes (Ho, 2009). Besides this, several institutions adopted capstone courses at the level of undergraduate education. Taiwan's universities and colleges are also encouraged to put emphasis on the development of curriculum maps to help learners select core and elective courses into order to cultivate their core and professional competencies.

All in all, in the upcoming 2011 institutional accreditation, 81 institutions have been requested to set up a set of generic attributes and core competencies for graduates and to explain how the intended learning outcomes can be achieved. Most important of all, the institutions have to establish a reliable assessment system in order to provide the relevant evidence for the exercise. In the second cycle of programme accreditation, student learning outcomes will be also expected to be embedded into each standard. Though HEEACT is still facing many big challenges, such as reviewers' training, communication with programmes and institutions, and faculty participation, it is expected to assist Taiwan's institutions in enhancing educational quality and developing graduates' global competitiveness.

Conclusion

To conclude, Taiwan's accrediting agencies are transforming the traditional accreditation model into a new learning outcomes based model, which has led Taiwan's institutions to develop diversified strategies to achieve the learning goals. According to the meta evaluation of HEEACT's accreditation, in the first cycle of programme based accreditation, three major goals have been achieved, including understanding the current situation of the overall quality of Taiwan's universities and colleges; driving universities and colleges to set up an internal self-enhancement mechanism; and assisting them in developing their own features and perusing excellence (Chan, 2010). To a certain extent, the quality assurance system has been moving Taiwan higher education into an 'era of quality'. The major concern is how to help higher education institutions implement the new approaches successfully on different campuses based on the mutual trust and understanding between HEEACT and institutions. Hopbach (2009) indicated that, "using learning outcomes is not only a challenge for higher education institutions in designing curricula and assessing students but also for quality assurance be it internal or external quality assurance" (p. 24) . Therefore, the crucial job is to develop close cooperation between accreditors and institutions; "accreditors must work collaboratively with higher education institutions to develop a common language that can explained the diverse approaches to addressing student learning outcomes." (Hawkins, 2009, p. 36).

References

Association of American Colleges & Universities (2008). ***New leadership for student learning and accountability***. Washington, D.C. AACU.

Chan, Y. & Yang, G. S. (2010). ***Report of meta-evaluation on the HEEACT's program accreditation from 2007 to 2009***. Taipei: Higher Education Evaluation & Accreditation Council of Taiwan.

CHEA (2001). ***Accreditation and student learning outcomes: A proposed point of departure.*** CHEA Occasional Paper, Washington, D.C. CHEA.

Ewell, P. (2008). ***U.S. accreditation and the future of quality assurance.*** Washington, D.C. CHEA Forum on the Collaborations among University Evaluation Agencies (Nov. 26, 2008).

Frye, R. (2009). ***Assessment, accountability, and student learning outcomes.*** Retrieved 27 April, 2009, from http://www.ac.wwu.edu/~dialogue/issue2.html

Hawkins, D., Lake, D., & Nielson, D., Tierney, M. (Eds.) (2006). ***Delegation and agency in international organizations.*** Cambridge, UK. Cambridge University Press.

Hawkins, J. (2009). ***“Internal and external quality assurance: implication for learning and accreditation: some observations from the University of California”***, Proceedings of Conference on QA and Student Learning Outcomes of HE in Asia-Pacific Region, Taiwan, pp. 55-72.

Higher Education Evaluation & Accreditation Council of Taiwan (2008a). ***2007 HEEACT annual report.*** Taipei: Higher Education Evaluation & Accreditation Council of Taiwan.

Higher Education Evaluation & Accreditation Council of Taiwan (2008b). ***HEEACT Handbook***. Taipei: Higher Education Evaluation & Accreditation Council of Taiwan.

Higher Education Evaluation & Accreditation Council of Taiwan (2009). ***2008 HEEACT annual report***. Taipei: Higher Education Evaluation & Accreditation Council of Taiwan.

Higher Education Evaluation & Accreditation Council of Taiwan (2010). ***2011 HEEACT’s Accreditation Handbook***. Taipei: Higher Education Evaluation & Accreditation Council of Taiwan.

Ho, S. H. (2009). ***‘Learning outcomes and quality assurance in Soochow University, Taiwan’***, Proceedings of Conference on QA and Student Learning Outcomes of HE in Asia-Pacific Region, Taiwan, pp. 93-108.

Hou A.Y.C. (2009). ***‘Student learning outcomes and quality assurance of higher education in Taiwan’***, Proceedings of Conference on QA and Student Learning Outcomes of HE in Asia-Pacific Region, Taiwan, pp. 55-72.

Marginson, S. (2007). The public/private divide in higher education: a global revision. ***Higher Education, 53 (3), pp. 307-333.***

Ministry of Education (2008). ***Introduction to higher education***. Taipei: Ministry of Education.

Ministry of Education (2009a). Promoting student quality in postsecondary education program. ***E-Journal***. Retrieved 27 April 2009 from http://72.14.235/search?q=cache:oOuGPd1NpH41:140.111.34.116/e9617-epaper/news.a

Ministry of Education(2009b) ***Teaching excellence program for colleges of technology and technical colleges.*** Retrieved from 27 April, 2009, http://pote.oit.edu.tw/rules/r9.pdf

Taiwan Assessment and Evaluation Association (2008). ***Evaluation reports.*** Retrieved Dec. 7, 2008, from http://www.twaea.org.tw/about.htm.

Technological & Vocational Educational Newsletter (2007, Oct.). ***“Establishing an evaluation mechanism of impartiality—an interview with President of National Yunlin University of Science & Technology”***, No. 176, Taipei: Ministry of Education.

UNESCO (2006). ***UNESCO-APQN Toolkit: Regulating the quality of cross-border education.*** Bangkok: UNESCO Asia and Pacific Regional Bureau for Education.

Wang, B. J. (2008). How is graduate performance evaluated? ***Evaluation Bimonthly, 15, pp. 24-25.***

Chait, R.(2002). The “academic revolution” revisited. In S. Brint (Ed.), ***The city of intellect.*** Stanford, California: Stanford University Press, pp. 293-321.

Schmidtlein, F. A., & Berdahl, R. O. (2005). Autonomy and accountability: Who controls Academe? In P. G. Altbach et al. (Eds.), ***American higher education in the twenty-first century, 2nd ed*** Baltimore: The Johns Hopkins University Press, ***pp. 71-90.***

Wollf, R. (2009). Future directions for U.S. higher education accreditation. In Terance W. Bigalke & Deane E. Neubruer (Eds). ***Higher Education in Asia/ Pacific, pp. 79-98.*** New York: Palgrave Macmillian.

Woodhouse, D. (June, 2010). The pursuit of international standards. ***International Leadership Colloquium.*** Madrid.

9

Speaking test anxiety among Korean university students

Hyun-Ju Kim and H. Douglas Sewell

Abstract

Test anxiety can have a great effect on one's test performance and results. In few places are the potentially negative consequences of such test anxiety a greater concern than in highly competitive situations where high stakes tests are all too common, as in S Korea. Having observed numerous candidates perform badly in tests due to anxiety, there is a need to understand the magnitude and nature of this issue with relation to the testing of English speaking abilities in the Korean context. In an effort to develop this understanding, two differing speaking test formats, a face-to-face interview format and a group interview format, were utilised. By using such formats we were able to investigate the relationship between such formats and candidate anxiety, as well as examine how the levels of this anxiety affected speaking test performance in and between individual interview and group test formats.

1. Literature Review

Test anxiety is one of the more important variables that affect test scores and has also been considered a predictor of exam performance. Liebert and Morris (1967) argued that test anxiety has the two dimensions of worry and emotionality. Worry is one cognitive dimension of considering the consequence of failure on a test, while emotionality relates to the autonomous nervous system and is seen to be evoked by evaluative stress.

Test anxiety is often measured by the Test Anxiety Inventory (TAI) (Spielberger, 1980). Trait anxiety differs from test anxiety in that the former is a more stable personal characteristic whereas the latter is more situation-specific. Thus levels of test anxiety can appear differently depending on contexts with respect to worry and/or emotionality.

Test anxiety only appears during the state of the test due to the fear of evaluation. Sarason and Sarason (1990) stated that more than 25% of American students performed worse, due to test anxiety, than their real ability. Other researchers such as Alpert & Haber (1960) and Tobias (1985) have however argued that facilitative anxiety might also work to improve performance.

2. Methods

2.1 Research Overview

The goal of this research project was to investigate test anxiety in Korean university level students in the two conditions of an individual interview speaking test and a group speaking test. In doing so, this research aimed to better understand the effect of speaking test format with respect to subjects' test anxiety and test performance.

2.2 Subjects

Subjects for this research were 47 university level students studying at a 4-year university outside Seoul. Subjects ranged from first to fourth year undergraduates and from a variety of majors. The gender ratio was approximately 40/60 for male/female with ages generally ranging from 20 to 27 years old.

All subjects were enrolled in optional English conversation courses led by one of the researchers on this project. As these classes were not level tested, subjects' levels were quite varied, ranging from high beginner to low advanced. All subjects were informed that participation was purely optional. Of over 50 subjects available for research, one chose not to opt in while a number of others were not able to participate due to timetable scheduling issues. During the speaking tests, one of the researchers in this study both conducted the speaking tests and rated subjects performance at the same time, while the other researcher observed the speaking tests.

2.3 Test Formats

As noted above, two speaking test formats were considered in this research. The first format was a one-on-one interview format lasting 9 - 9.5 minutes. This test included a two way conversation with the examiner as well as preparation and speaking time for a short monologue by the candidate. The second test format was a group speaking test in which 3-4 subjects were given a list of themed conversation questions to prompt discussion. As most students of English in Korea have previously experienced one-on-one speaking tests, and similar themed group conversation topics, subjects were thus felt to have had approximately equal overall exposure to both formats used in this research.

Both speaking tests were marked using the same specially developed rubric. This rubric contained nine levels, corresponding to low beginner through high advanced, and measured linguistic output in terms of four categories: fluency, language use, grammar, and pronunciation. This rubric provided at least a reasonable measure of the language produced in each of the test formats.

2.4 Research Instruments

Three survey instruments were used in this research, with the first two given before either of the speaking tests and the third after the second speaking test. The first instrument given was the Foreign Language Classroom Anxiety Scale (FLCAS) (Horwitz 1986) (Appendix A). The FLCAS was translated into Korean by a professional translator and then reviewed by a second native Korean speaker familiar with the field of English education in Korea. Trialling resulted in minor alterations to the Korean translation before the final version was administered. The second survey instrument was a modified version of the State-Trait Anxiety Index (Speilberger 1983 in Baker 2011) (Appendix B). As with the FLCAS, the instrument was translated, trialled and edited before the final version was used.

The final survey instrument was a six-item questionnaire written specifically for this research project (Appendix C). It was observed that the more anxious subjects appeared, the more they felt they did not perform to what seemed to be their real ability. Based on these observations, questions 1 and 3 of this survey asked subjects to rate on a seven point scale how well they felt they performed on the group and individual speaking tests compared to their overall perception of their actual English ability. Questions 2 and 4 then asked subjects to give reasons for their answers to the respective previous question. Using another seven point scale, question 5 asked subjects to indicate which of the two tests they found to be more difficult. Question 6 then asked subjects to give reasons for their previous answer.

2.5 Data Analysis

Quantitative data from all three survey instruments was entered into SPSS (Ver. 18) for data analysis while qualitative results from the final survey questions 2, 4 and 6 were translated into English with salient comments noted for further consideration. The video recordings of the individual and group interviews were viewed with notes of interesting points taken and video excerpts produced as needed.

3. Findings and Discussion

3.1 Speaking Performance Between the Two Formats

The first aspect of the data to be considered was an examination of the means and standard deviations of the results of the two speaking tests. These are presented in Table 1 below.

Rating Criteria	Interview Formats	Number	Mean	SD
Fluency	Individual	47	5.53	1.248
	Group	47	6.12	.991
Language Use	Individual	47	5.89	1.107
	Group	47	6.17	.985
Grammar	Individual	47	5.72	1.117
	Group	47	6.04	1.041
Pronunciation	Individual	46	5.82	1.216
	Group	46	6.26	1.042
Composite	Individual	47	5.69	1.086
	Group	47	6.01	.975

Table 1 - Mean and Standard Deviation of the Two Interview Formats.

As highlighted in Table 2 below, the correlation coefficient between composite scores of the two interviews was .809 (P=.000). This indicates that the means of the two test formats are generally correlated. In order to compare the mean difference between the two different interview formats, a paired sample t-test was employed. The results of the t-test are presented in Table 3. Statistically these results indicate that across all the rating criteria, there was a significant difference in test-takers' speaking performance between the two interview formats, with the test scores of the four analytic rating criteria and composite score in the group format higher than in the individual interviews.

	Number	Correlation Coefficient	P-value
I_Fluency-G_Fluency	47	.646	.000
I_LU-G_LU	47	.674	.000
I_Gram-G_Gram	47	.739	.000
I_Pron-G_Pron	47	.720	.000
I_Comp-G_Comp	47	.809	.000

Note: I = individual interviews; G = group interview; LU = language use; Gram = grammar; Pron = pronunciation; Comp = composite scores

Table 2 - Paired Correlation Coefficient.

	Mean	*SD*	*t*	*df*	*p*
I_Fluency-G_Fluency	-.59	.970	-4.209	46	.000
I_LU-G_LU	-.27	.852	-2.225	46	.031
I_Gram-G_Gram	-.31	.783	-2.794	46	.008
I_Pron-G_Pron	-.43	.860	-3.428	45	.001
I_Comp-G_Comp	-.31	.646	-3.386	46	.001

Table 3 - Paired T-test Between the Two Interview Formats.

3.2 Foreign Language Anxiety and Test Anxiety

Subjects' foreign language anxiety and test anxiety were measured by translated versions of the Foreign Language Classroom Anxiety Scale (Horwitz, 1986) and the State-Trait Anxiety Index (Speilberger, 1983). Through factor analysis as shown in Table 4 below, it was found that subjects' foreign language anxiety could be categorised into the four factors of general uneasiness, impatience, awareness of others, and fear of teachers, though not with a strong overall result. Test anxiety was in the same way found to be able to be categorised into the three factors of; discontent, worry, and tension. Again these results were not strong overall, although items 8 ('I feel unsatisfied with this test'; m=4.63) and 11('I feel unconfident with myself for this test'; m=4.30) did show interesting results and suggested that subjects' self-confidence on the test was on the low side. These results lead to the second research question focussing on the nature of the relationship between test anxiety and speaking performance in this research.

Foreign Language Anxiety	Items	Mean	*SD*
General Uneasiness	1,10, 16, 17, 19, 21, 30	2.42	.707
Impatience	3, 12, 20, 26, 27	2.80	.985
Awareness of others	9, 24, 31, 33	2.97	.561
Fear of teachers	4, 15, 29	2.64	.864

Table 4 - Foreign Language Anxiety.

Test Anxiety	**Items**	**Mean**	***SD***
Discontent	1, 3, 9, 12, 15	3.29	.191
Impatience	7, 8, 11, 14, 18	3.72	1.42
Awareness of others	2, 6	3.47	1.532

Table 5 - Test Anxiety.

3.3 Test Anxiety and Speaking Performance

As seen in Table 6 below, when considering speaking performances within the individual and group speaking tests with respect to test anxiety levels, no significant difference within the formats was found. This result implies that test anxiety levels in different speaking test formats do not affect subjects' speaking performance significantly. Therefore, the assumption that anxiety would inhibit production of a second language was not statistically supported. Although many researchers have suggested the possibility that foreign language anxiety and test anxiety interfere with language learning and language performance, surprisingly those who had higher levels of test anxiety in this research received scores similar to those who had lower levels.

			SS	*df*	*MS*	*F*	*P*
Individual Interview	Fluency	Between groups	.298	2	.149	.090	.914
		Within groups	71.115	43	1.654		
	Language Use	Between groups	1.544	2	.772	.604	.551
		Within groups	54.913	43	1.277		
	Grammar	Between groups	.705	2	.353	.268	.766
		Within groups	56.621	43	1.317		
	Pronunciation	Between groups	.880	2	.440	.291	.749
		Within groups	65.055	43	1.513		
	Composite	Between groups	.656	2	.328	.263	.770
		Within groups	53.583	43	1.246		
Group Interview	Fluency	Between groups	3.016	2	1.508	1.565	.221
		Within groups	41.441	43	.964		
	Language Use	Between groups	2.630	2	1.315	1.347	.271
		Within groups	41.979	43	.976		
	Grammar	Between groups	2.306	2	1.153	1.062	.355
		Within groups	46.672	43	1.085		
	Pronunciation	Between groups	.702	2	.351	.310	.735
		Within groups	47.609	43	1.134		
	Composite	Between groups	1.759	2	.880	.906	.412
		Within groups	41.741	43	.971		

Table 6. - One-Way Analysis of Variance Summary for Speaking Performances.

However, the results of subjects' perceptions of test anxiety of the two different test formats showed that there was a relationship between test anxiety and test formats, and that relationship was related to different levels of test performance. Table 7 below shows that there were significant differences in subjects' perceptions of the level of test anxiety between the two different interview formats. This indicates that subjects showed higher anxiety during the individual interview test compared to the group test and perceived that they performed closer to their real language ability in the group speaking test compared to the individual interview test.

	Mean	***SD***	***t***	***df***	***p***
C1_C3	-.54	.780	-4.723	45	.000

Table 7 - Paired T-test of Subject Perception Between the Two Interview Formats.

Highly illustrative of these results are the answers to the open-ended questions on the last survey instrument. These results, translated and presented below, highlight the relationship between test format and anxiety, with subjects suggesting that they felt more comfortable in the group speaking test compared to the individual interview test.

> *"I was very nervous, but I guess I performed my language ability well. Thanks to other group members, I could reduce my nervousness."*
>
> *"I felt more comfortable in the group test because I was being with other students."*
>
> *"I was too nervous to show my speaking ability well. Being only with the professor (interviewer) made me very nervous."*

Subjects also suggested that they performed better in the group speaking test since they felt less test anxiety than in the individual interview test.

> *"Individual interview was more difficult since I had to keep talking."*
>
> *"It was very difficult to perform my speaking ability in the individual interview test while looking at the professor who was judging my English ability. Group test was more comfortable so I think I performed better in it."*
>
> *"I felt like the speaking errors were more easily shown in an individual test, so I guess I performed better in group tests."*
>
> *"Group test was better for me since I could get some help from the other students during the test."*

4. Conclusion

The results of this research give significant insights into Korean test-takers with respect to speaking test formats, test anxiety within these formats and their performance, both statistically and perceptually, across these formats. The results support some findings from previous research on test anxiety, but limitations in this research, including the need for a larger sample size, warrant caution in the generalisation of these findings and suggest that further research is necessary.

With such limitations addressed, it may be possible to clarify the ways in which different levels of test anxiety across different speaking test formats effects test performance.

Overall, however, this research cautiously suggests that group speaking tests rather than individual face-to-face interview tests may be more suitable in some speaking test situations, such as in university conversation practice classes in the Korean context, as such group speaking tests may allow subjects to demonstrate their actual speaking ability more fully.

References

Alpert, R. & Haber, R.N. (1960). Anxiety in academic achievement situations. ***Journal of Abnormal and Social Psychology. 10, pp. 207-215.***

Baker, C., Clark, D. M., Pertaub, D., & Slater, M. (2011). ***State Trait Anxiety Inventory.*** http://www.cs.ucl.ac.uk/research/vr/Projects/SocialPhobias/exp2000/docs/questionnaires/stai.doc (Accessed June 5th, 2011).

Horwitz, E., Horwitz, M. & Cope, J. (1986). Foreign language classroom anxiety. Modern Language Journal. ***70, pp. 125-32.***

Liebert, R. M. & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: a distinction and dome initial data. Psychological Reports. ***20, pp. 975-978.***

Sarason, I. G. (1961). Test anxiety and the intellectual performance of college students. Journal of Educational Psychology. ***52 (4), pp. 201-206.***

Sarason, I. G., & Sarason, B. R. (1990). Test anxiety. In H. Leitenberg (Ed.), Handbook of Social and Evaluative Anxiety, ***pp. 475-496***. New York: Plenum Press.

Spielberger, C. D. (1980). ***Test Anxiety Inventory. Preliminary Professional Manual.*** Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). ***Manual for the State-Trait Anxiety Inventory.*** Palo Alto, CA: Consulting Psychologists Press.

Tobias, S. (1985). Test anxiety: interference, defective skills, and cognitive capacity. Educational Psychologist. ***20, pp. 135–142***

Appendix A - Survey 1a - English Version

Language Test Format Research - Survey 1a

Please indicate the extent you agree or disagree with each statement by circling the number which best expresses your feelings for each statement. When considering your answers, please think of yourself in general, not in relation to any specific class you have ever taken.

		Disagree	Disagree Strongly	Neutral	Agree	Strongly Agree
1	I never feel quite sure of myself when I am speaking in my foreign language class.	1	2	3	4	5
2	I don't worry about making mistakes in language class.	1	2	3	4	5
3	I tremble when I know that I'm going to be called on in language class.	1	2	3	4	5
4	It frightens me when I don't understand what the teacher is saying in the foreign language.	1	2	3	4	5
5	It wouldn't bother me at all to take more foreign language classes.	1	2	3	4	5
6	During language class, I find myself thinking about things that have nothing to do with the course.	1	2	3	4	5
7	I keep thinking that the other students are better at languages than I am.	1	2	3	4	5
8	I am usually at ease during tests in my language class.	1	2	3	4	5
9	I start to panic when I have to speak without preparation in language class.	1	2	3	4	5
10	I worry about the consequences of failing my foreign language class.	1	2	3	4	5
11	I don't understand why some people get so upset over foreign language classes.	1	2	3	4	5
12	In language class, I can get so nervous I forget things I know.	1	2	3	4	5
13	It embarrasses me to volunteer answers in my language class.	1	2	3	4	5
14	I would not be nervous speaking the foreign language with native speakers.	1	2	3	4	5
15	I get upset when I don't understand what the teacher is correcting.	1	2	3	4	5
16	Even if I am well prepared for language class, I feel anxious about it.	1	2	3	4	5
17	I often feel like not going to my language class.	1	2	3	4	5

		Disagree	Disagree Strongly	Neutral	Agree	Strongly Agree
18	I feel confident when I speak in foreign language class.	1	2	3	4	5
19	I am afraid that my language teacher is ready to correct every mistake I make.	1	2	3	4	5
20	I can feel my heart pounding when I'm going to be called on in language class.	1	2	3	4	5
21	The more I study for a language test, the more confused I get.	1	2	3	4	5
22	I don't feel pressure to prepare very well for language class.	1	2	3	4	5
23	I always feel that the other students speak the foreign language better than I do.	1	2	3	4	5
24	I feel very self-conscious about speaking the foreign language in front of other students.	1	2	3	4	5
25	Language class moves so quickly I worry about getting left behind.	1	2	3	4	5
26	I feel more tense and nervous in language class than in my other classes.	1	2	3	4	5
27	I get nervous and confused when 1 am speaking in my language class.	1	2	3	4	5
28	When I'm on my way to language class, I feel very sure and relaxed.	1	2	3	4	5
29	I get nervous when I don't understand every word the language teacher says.	1	2	3	4	5
30	I feel overwhelmed by the number of rules you have to learn to speak a foreign language.	1	2	3	4	5
31	I am afraid that the other students will laugh at me when I speak the foreign language.	1	2	3	4	5
32	I would probably feel comfortable around native speakers of the foreign language.	1	2	3	4	5
33	I get nervous when the language teacher asks questions which I haven't prepared in advance.	1	2	3	4	5

Appendix B - Survey 1b - English Version

Language Test Format Research - Survey 1b

Please indicate the extent you agree or disagree with each statement by circling the number which best expresses your feelings for each statement.

1: I feel pleasant								
Almost never	1	2	3	4	5	6	7	Almost always
2: I feel nervous and restless								
Almost never	1	2	3	4	5	6	7	Almost always
3: I feel satisfied with myself								
Almost never	1	2	3	4	5	6	7	Almost always
4: I wish I could be as happy as others seem to be								
Almost never	1	2	3	4	5	6	7	Almost always
5: I feel rested								
Almost never	1	2	3	4	5	6	7	Almost always
6: I am 'calm, cool and collected'								
Almost never	1	2	3	4	5	6	7	Almost always
7: I feel that difficulties are piling up so that I cannot overcome them								
Almost never	1	2	3	4	5	6	7	Almost always
8: I worry too much over something that doesn't really matter								
Almost never	1	2	3	4	5	6	7	Almost always
9: I am happy								
Almost never	1	2	3	4	5	6	7	Almost always
10: I have disturbing thoughts								
Almost never	1	2	3	4	5	6	7	Almost always
11: I lack self-confidence								
Almost never	1	2	3	4	5	6	7	Almost always
12: I feel secure								
Almost never	1	2	3	4	5	6	7	Almost always
13: I make decisions easily								
Almost never	1	2	3	4	5	6	7	Almost always
14: I feel inadequate								
Almost never	1	2	3	4	5	6	7	Almost always
15: I am content								
Almost never	1	2	3	4	5	6	7	Almost always
16: Unimportant thoughts run through my mind and bother me								
Almost never	1	2	3	4	5	6	7	Almost always

17: I take disappointments to heart and I can't put them out of my mind		
Almost never	1 2 3 4 5 6 7	Almost always
18: I get in a state of tension or turmoil when I think about my recent concerns and interests		
Almost never	1 2 3 4 5 6 7	Almost always

Appendix C - Survey 2 - English Version

Language Test Format Research - Survey 2

1. Please circle the statement which best describes your feelings after the GROUP speaking test.

On the GROUP test, I feel I performed:

far above my true ability	somewhat above my true ability	a bit above my true ability	at my true ability	a bit below my true ability	somewhat below my true ability	far below my true ability

2. If you performed above or below what you felt was your true ability, please explain why you felt you performed above/below your true ability? You may write in Korea.

____

____

3. Please circle the statement which best describes your feelings after the INDIVIDUAL speaking test.

On the INDIVIDUAL test, I feel I performed:

far above my true ability	somewhat above my true ability	a bit above my true ability	at my true ability	a bit below my true ability	somewhat below my true ability	far below my true ability

4. If you performed above or below what you felt was your true ability, please explain why you felt you performed above/below your true ability? You may write in Korea.

____

____

5. Please make a circle around the statement which you feel best describes which test felt harder overall, the GROUP test or the INDIVIDUAL test?

The GROUP test was much harder	The GROUP test was somewhat harder	The GROUP test was a bit harder	Both tests were about equally hard	The individual test was a bit harder	The individual test was somewhat harder	The individual test was much harder

6. Please let me know why you felt one test was harder than the other. You may write in Korean.

____

____

THE TEACHERS’ PERSPECTIVE

10

Learning from the experts: A comparative study between expert and novice teachers in assessing ESL writing

Clarence Jerry, Moses Samuel and Jariah Mohd. Jan

Abstract

Writing is a productive skill that is highly prioritised in English as a Second Language (ESL) classrooms. Teachers are expected not only to teach the skill well but at the same time should possess essential competence in assessing it. This case study investigates the ESL teachers' practices in assessing students' written work. Interviews were conducted to gather their views. The findings revealed that expert teachers engaged in multiple reviewing activities focus more on meaning-related concern of an essay. Novice teachers, however, emphasised students' language errors. These findings provided insights on the need to assist novice teachers in improving their assessment practices.

Introduction

Teachers' written feedback is an essential aspect in any English language writing course. In studies that have examined it, feedback is also associated clearly with writing improvement, especially for older students, and students appreciate it (Ferris, 1997). According to Graves (1983), teacher feedback and the opportunity to revise written work based on this feedback are key to students' development as writers. The way a teacher provides feedback will have direct impact on whether students become successful or unsuccessful writers. Apart from that, novice writers need guidance to evaluate, modify, or restructure their ideas and to add and delete content to improve their writing (Condon, 2009; Keppner, 1991; Olson & Raffeld, 1987).

Studies done on writing suggest that feedback plays a central role in increasing the learner's achievement. The more information learners have about their writing, the better they understand how to improve performance (Cardelle & Corno, 1981). Learners of writing need feedback, not only to monitor their own progress, but also to take others' views and adapt a message accordingly (Flower, 1979). An additional effect of corrective feedback may be the enhancement of learners' metalinguistic awareness (Swain, 1995), an important step in their appropriation of the written system. According to Nelson and Schunn (2009), although providing feedback is commonly practised in education, there is no general agreement regarding what type of feedback is most helpful and why it is helpful. In addition, there is little research which has focused on the comparison between the expert and novice teachers' practices in assessing students' written work and how the expert teachers' way of assessing differs from those of novice teachers.

The Study

This study aims to investigate the differences between expert and novice teachers in assessing students' written work.

Methodology

A qualitative case study research design was employed. Based on purposive sampling, there were eight participants (four experts and four novices) in this study. The expert raters were experienced teachers with more than 10 years of teaching experience while the novice raters had less than 3 years of teaching experience. The data collection procedures involved interviews. The raters or teachers were requested to give their views after marking two given essays. Their views were video recorded. The interview sessions were transcribed verbatim and analysed by identifying emerging themes and categories of issues. The coded data were cross-checked with the participants to enhance reliability.

Results and Discussions

Expert Raters

(i) ER 1

ER1 employed two widely-used knowledge states in assessing both sample essays - the choice of expressions and words. In the exchange below, ER 1 provided the following related information:

> ***Example 1***
> *...a B (essay) to me would mean that the command of the language is there but the "sophistication" of it is not there. So sophistication...I would put it as complex structures and vocabulary that is precise.*
>
> (ER1/VP/Transcript/13.6.07)

As an expert rater, ER 1 was very thorough in his assessment. His first main priority was to get an overall glimpse of what the writing was all about.

> ***Example 2***
> *Normally, if you give me some essays to look at, let say three...I'll read ... definitely the first paragraph first. That's the first thing I will do. I don't need to read the rest of the paragraphs...I just need to read one. So, when I read it, basically...what I'm looking for...like I said earlier...the ability of the candidate to communicate to me....*
>
> (ER1/VP/Transcript/13.6.07)

(ii) ER 2
ER 2 stressed the choice of expression in both essays. During the interview, when asked to comment on how she would usually assess her students' writing, ER 2 divided her comments into three categories: structure, content and language. This can be seen from the interview with ER 2 in Example 3.

> ***Example 3***
> *Actually, when I mark an essay, I will look at the criteria. Generally I will look at three perspectives. One is structure, the other one is content and the other one is the language itself.*
>
> ER2/Interview/Comments/14.6.07

As for organisation, she would be looking at the main points (topic sentences) and how they develop into paragraphs, which would give her an impression of the student's competence, as clearly shown in Example 4:

> ***Example 4***
> *I will be looking at organisation, the details...how good the details...how good are the details. Okay, so we are looking at main point...elaboration...the organisation of ideas itself. How mature the students are...and so on.*
>
> ER2/Interview/Comments/14.6.07

(iii) ER 3
ER 3 did not really go into details on how he would assess students' writing. Obviously, what he generally did was skim through the writing to determine what it was all about before awarding marks based on holistic grading. He also focused his assessment based on the students' command of grammar, sentence structure and vocabulary, as illustrated in Example 5:

> ***Example 5***
> *Generally in marking an essay question, I will first skim through the essay because I don't want to find myself penalising the kids...the students for the grammatical (errors), the vocabulary and the error...the semantics...so what I'll do is I'll quickly read through to see whether I understand what the candidate is writing about and where I would put them in the grade... I would then read the essay and see whether I can give the student a better grade from the general grade by looking at the grammar, sentence structure and err...vocab.*
>
> ER3/Interview/Comments/14.6.06

(iv) ER 4
ER 4 produced many more comments as compared to the other expert raters. When given the task of assessing writing, ER 4 showed that he placed his priority on meaning and language. This is pretty obvious from his interview response in Example 6 when asked on how he would mark a piece of writing.

> ***Example 6***
> *...I'm actually looking at basically at two things. The first is I'm looking at for the ability to communicate their ideas to us...the second is command of the language. What do I actually look for when I say the ability to communicate...I'm actually looking at whether or not they can convey their ideas to me in English. Number two...when I read it, do I have to translate it from another language, is there a need to paraphrase the thing or I can say that it's not something I would recognise but I would need to do a bit of work. So, I'm looking for what I would call...I'll use the phrase as in it comes clearly to me without any element*

> *of doubt. I do not want to see things that I need to infer or make a decision.*
>
> ER3/Interview/Comments/14.6.06

Overall, ER 4 showed concern for language use, content, and development of the essays.

Novice Raters

(i) NR 1

NR 1 would skim through each piece of writing given to him and decide if it was a good piece of writing. After skimming through, he would scan through the paragraphs for clarity of ideas and identify errors. Though he would try to identify all the errors in that particular piece of writing, the focus was more on the meaning (Example 7). He used this approach to decide if errors committed had impeded meaning in any way. Overall, he would focus on the clarity of ideas (meaning) first, then proceed to aspects of grammar and spelling errors.

> ***Example 7***
>
> *Normally I would just read through once and get the tips but sometimes err, if it's a good piece of writing, meaning err.... Good in the sense of I can understand everything the writer is trying to say without really struggling. Then I would go paragraph by paragraph err... because I know I can... I can get the meanings straight away without having to read so many times. So... I will read (a) paragraph, get the whole meaning of the-that paragraph, and I will identify whether there's grammar error(s) or spelling errors...*
>
> NR1/Interview/Comments/15.6.06

(ii) NR 2

NR 2 provided comments mostly on sentence level errors. She underlined and wrote comments below the text, circled, crossed out, and used codes such as N for noun and T for tense, to indicate the category of the errors. Apart from that, she placed emphasis on the meaning aspect for every paragraph. In the exchange below, NR 2 provided the following information:

> ***Example 8***
>
> *Basically, when I mark students' writing I would first mark all the errors. At the same time I will try to see if the errors impede meaning. By identifying all the possible errors, it would help me decide if the essay conveys the appropriate meaning.*
>
> NR2/Interview/Comments/15.6.06

(iii) NR 3

NR 3 would skim through the essays for general understanding. After that, she would scan through the paragraphs for errors during the verbal protocol analysis. She would cross out or underline errors, and put ^ marks indicating that some words needed to be added to make the sentence correct. This was also indicated in her interview response when asked how she would mark an essay as given in Example 9.

> ***Example 9***
>
> *When I get a pile of paper to mark, normally, I'll choose the best students...I like to read the good ones first. When I get to the paper, what I'll do is I'll read the paper from beginning to the end first...then after that...I go from paragraph to paragraph. I'll look for whatever errors, whatever points I need to...whatever marks I need to give.*
>
> NR3/Interview/Comments/25.4.07

(iv) NR 4

NR 4 seemed to be very confident in assessing the two pieces of writing during the VPA. He skimmed through the first and last paragraph first to get an overall idea of what the essay was all about. He used an holistic approach of assessment to determine students' writing proficiency level as illustrated in the discourse as in Example 10:

> ***Example 10***
>
> *Usually what I will do before I mark a test paper especially in writing is that I will read first paragraph and I will read the last paragraph. Because generally I think that if you read the first paragraph and the last paragraph, you get the whole idea of the direction of the story first and ah...secondly, by doing so...you are able to give a holistic view on student's level of the English language and more or less set the benchmark in my mind as to how good this student will be and how good I hope he will be actually.*
>
> NR4/Interview/Comments/15.6.06

NR 4 also scanned through the writing to look for a storyline which would show how effectively students had developed the ideas presented in the paragraphs.

> ***Example 11***
>
> *... I would scan the whole story as I go along and ah...what I look for is story line. I think it is very important that students are able to develop their ideas.*
>
> NR4/Interview/Comments/15.6.06

From the qualitative data, it can be noted that the novice raters in this study seemed to be grammar focused and they identified most of the surface errors prior to making any decision on the content. On the other hand, the expert raters were more oriented toward content and meaning-focused assessment. Specifically, the expert raters tended to conceptualise assessing activity as a macro-strategy similar to the findings by Eckes (2008), who discovers that expert raters tend to have enough knowledge to articulate their reviewing process of a written work. The expert raters did not overlook grammatical errors but they seemed to focus on errors that inhibited communication. On the other hand, novice raters may focus more on surface-level errors because they are the easiest to detect and respond to. However, comments on content require a higher degree of judgement and most likely take more time and so they were attended to less frequently or in less detail. Leki (1991), in fact, speculates that because errors in grammar and mechanics are more concrete than meaning-related problems, they are relatively easier to correct.

The findings also indicate that expert raters engaged in multiple reviewing activities during assessment of writing, including many revisions that were not concerned with simple matters of surface accuracy. Cumming, Kantor and Powers (2002) support this by revealing in their study that ESL and EFL expert raters attended more extensively to language, rhetoric and ideas as a whole rather than focusing on a specific element as in the case of novice raters. This pattern seemed to be directly attributable to the kind of assessing experience which the expert raters seemed to have accumulated over their years of classroom practice and marking public examination papers.

Implications and Recommendations

Based on the study, it can be concluded that ELT teachers need to focus on meaning-related concerns in a piece of written work, in other words the content of an essay, before focusing on sentence-level language errors. Teachers need to respond to content and organisation before attending to grammatical errors. This will avoid premature editing and making revisions to a text at a surface level instead of at global level. At the same time, the close focus on features of language use which accompanies the discussion of learner performances serves a valuable professional development function. In order to encourage students to develop their own voices (in writing), teachers need to play their role by regularly making conscious choices during the assessing and giving feedback process. In planning professional development activities for teachers, teacher educators may find the model of the expert rater formulated in this study useful in making the right choice of training strategy. Students do not grow as writers, and teachers do not grow as instructors, in the absence of high-quality feedback. As with students, teachers need opportunities for collaborative assisted professional development in order to improve their practice.

Conclusion

This study has highlighted the differences in how expert raters and novice raters assess writing. Understanding the deep structures of knowledge, or schemata, allows the expert to see large and meaningful patterns in problem-solving. Emphasis is given by the expert raters to choice of expression and clarity in students' writing. It can be concluded that novice teachers need to prioritise their comments to focus on meaning-related issues (the content of an essay) before focusing on sentence-level language errors. Ultimately, the findings from this study could help institutions to plan a structured programme of training for novice teachers.

References

Cardelle, M. & Corno, L. (1981). Effects on second language learning of variations in written feedback on homework assignments. ***TESOL Quarterly, 15(3), pp. 251-261.***

Condon, W. (2009). Looking beyond judging and ranking: Writing assessment as a generative practice. ***Assessing Writing, 14(3), pp. 141-156.***

Cumming, A., Kantor, R. & Powers, D. E. (2002). Decision making while rating ESL/EFL Writing Tasks: A descriptive framework. ***The Modern Language Journal, 86(1), pp. 67-96.***

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. ***Language Testing, 25(2), pp. 155-185.***

Ferris, D. R. (1997). The influence of teacher commentary on student revision. ***TESOL Quarterly, 31, pp. 315-339.***

Flower, L. S. (1979). Writer-based prose: A cognitive basis for problems in writing. ***College English, 41(1), pp. 19-37***

Graves, D. (1983). ***Writing: Teachers and Children at Work.*** London: Heinemann Educational Books.

Keppner, C. (1991). An experiment in the relationship of types of written feedback to the development of second-language writing skills. ***Modern Language Journal, 75, pp. 305-313.***

Leki, I. (1991). The preference of ESL students for error correction in college-level writing class. ***Foreign Language Annals, 24, pp. 203-218.***

Nelson, M. M. & Schunn, D. S. (2009). The nature of feedback: how different types of peer feedback affect writing performance. ***Instructional Science, 37(4), pp. 375-401.***

Olson, M. W. & Raffeld, P. (1987). The effects of written comments on the quality of student compositions and the learning of content. ***Reading Psychology, 4, pp. 273-293.***

Swain, M. (1995). Three functions of output in second language learning. In Cook, G. and Seidlhofer, B. (Eds.), ***Principles and Practice in Applied Linguistics: Studies in honour of H. G. Widdowson, pp. 125-144.*** New York: Oxford University Press.

11

Classroom observation and teacher assessment: '8 Photographs'

John Hankinson

Abstract

This paper deals with experimental work carried out in three countries in the Gulf region, (Saudi Arabia, Bahrain and Qatar) between 2006 and 2011. The experiment drew neither on previously published research nor did it attempt to have specific significance for research purposes. The focus in each case was the observation of teachers in their classrooms, and the possibility of their developing self-assessment skills and reflective practices when reviewing their own teaching. A further consideration in devising the experiment was the sustainability of sound (self) assessment techniques at the local (school) level, once the official project ended.

Background

The experiment described in this paper is the result of my work between 2006 and the present, carried out in three Gulf states: Saudi Arabia, Bahrain, and Qatar. In each of the above states, my work was with primary and secondary teachers of English as a foreign language. The reason for this experiment resulted from a perceived need to encourage greater self-reflection among teachers, common to all three education systems (which had been investing heavily in education change and innovation).

These three countries provide sufficient commonality for a single experiment to be relevant to developmental requirements in each individual case. In the Kingdom of Saudi Arabia (KSA) I worked with a British Foreign Office (FCO) project, the remit of which was to look at textbook provision across the whole Kingdom, following curriculum reforms in the teaching of English as a foreign language. Essentially, the Kingdom required teachers to implement the new syllabus in state (government) schools at primary and secondary levels without the benefit of suitable new textbooks to match and support the new syllabus. The FCO Project attempted to address this problem by working with both educationalists and publishers to elaborate the new notional syllabus into a functional syllabus, integrating new textbooks and other materials into the fabric of the curriculum. The second project on which I worked, in Bahrain, was very similar, except that the project was further advanced, in that both a new curriculum for English as a foreign language had been agreed and new textbooks commissioned. However, due to budgetary constraints, the new textbooks were not custom-written for Bahrain but rather adapted from existing published materials by the provider. Starting from a textbook in use in the USA designed for non-native speakers of English learning English as a second language, (rather than as a foreign language), the Bahrain edition was created over a period of approximately three years, and the new books were phased into schools over a three year period. In the third project, in Qatar, the focus was slightly different, in that teachers of Science and Maths at primary level were being asked to teach their subject in English to children whose first language was Arabic. An attempt to introduce Content and language integrated learning (CLIL) methodology was thus being made.

The commonality between these projects is clearly in their developmental nature. In all three projects the need for teachers to develop new skills rapidly to undertake the challenges of curriculum change was central to the success of the changes being attempted. In all three projects relatively under-qualified and under-trained teachers were bearing the brunt of the need to adapt to a new curriculum and new methodology, together with the need to understand and adopt new and unfamiliar materials in a very short time span. Understandably, this caused both apprehension and insecurity, not always addressed by sufficient in-service training. Frequently, teachers were challenged by the situation in which they found themselves involved, and needed both reassurance and support to develop their skills, in order to cope with these changes.

The Model

In each of the first two projects, my role was twofold: consultant and mentor. Whilst involved in decisions at Ministry level, in addition I had a practical role as mentor, both at classroom level, working directly with teachers, and at the level of their advisors, advisory teachers and inspectors. In the third project I worked directly with teachers in a mentoring role, delivering formal training and informal mentoring, which was classroom based.

The overwhelming need of teachers in all three projects was to have their work examined in a non-judgmental way, in other words to have their work evaluated without that evaluation being seen as either an overt or covert assessment, in any formal sense. It was, therefore, difficult to apply the normal model of classroom-based evaluation, which is usually the result of a classroom (lesson) visit, followed by some sort of de-briefing and resulting in a written record being created by the assessor of the lesson visited.

The above model was in use in all of the above projects, but was the subject of suspicion and mistrust, in many cases. The contradiction between the need to improve teaching skills and those emerging skills being subjected to formal scrutiny resulted in many teachers reverting to the formulaic, tried and tested methodology which required no risk taking, rather than the bolder experimental approach, which the challenges of rapid and radical curricular change seemed to indicate.

A new model of classroom observation (the ‘8 Pictures’ approach) was developed, in order to engage teachers in their own (self) assessment, and to develop a culture of self-evaluation and of a more reflective approach to teaching. This technique was employed in working with individuals, post lesson observation, and in workshop situations, where materials from ‘real’ lessons could be analysed and discussed The technique involved replacing the traditional notebook or proforma with a digital camera. Throughout the lesson visit, numerous photographs were taken in an attempt to illustrate the lesson almost as one would do so using a video film. Stills were preferred to video, however, due to the inherent simplicity of taking unlimited digital photographs without any particular expertise, in contrast to the considerable technical expertise required to make even a short video sequence.

At the end of the lesson, the photographs were reviewed and edited by the observer, resulting in 8 being chosen, as representative of the lesson observed. The number was chosen to be representative of the standard lesson of around 45 minutes to one hour, and each picture chosen was selected on the basis of being representative of a theme within the lesson. These themes, (not an exhaustive list) included: ‘warm up’; ‘presentation’; ‘group work’; ‘working together’; ‘class dynamics’; ‘plenary’, etc. In each case, the photographs were collated immediately after the lesson, becoming the basis of post-lesson discussion between the teacher and observer (mentor).

Discussion stages

Discussion of the selected frames always began with an overview, and with an invitation to the teacher involved to comment generally on the sequence. The question, “How did you think the lesson went?”, (frequently heard in de-briefing sessions of this kind), was strictly avoided, and the emphasis was on a ‘stream of consciousness’ type approach, where the teacher having been observed was encouraged simply to look at the pictures and to comment in any way (s)he felt appropriate.

However, this initial stage was then moved forward into stage two, where the teacher was asked to focus on the language of the picture(s) individually, working through the 8 frames in sequence, and the focus was twofold:

a. What was said at that point by the teacher.
b. What was said at that point by the child(ren) in the frame.

Frequently, in order to facilitate this, the frame was printed, A4 size, and the teacher involved was asked to create a ‘speech bubble’, rather like a cartoon-strip, to record the ‘voices’ of the children in the frame, and to create a caption underneath the picture, to record the teacher’s ‘voice’.

This technique enabled teachers to focus very specifically on classroom language, an important focus for their personal professional development, especially as many were non-specialist (teachers of English), and more than a few were entirely untrained in English beyond their own experience of learning English as a subject at school or college level.

However, the main focus of the discussion was in terms of the content suggested by the frames. In each case, the teacher was encouraged to describe the scenario around the picture and to attempt an explanation of the intention and result of the lesson planning which had resulted in what could be seen in the frame. Once this had been done, the teacher was sometimes asked for clarification, followed by an invitation to write a ‘report’ on the lesson. This report was basically a paragraph written to describe each frame individually. Once the technique had been adopted and used regularly this report included not only narrative but also informed comment on the frame’s significance within the overall scope of the lesson, both in terms of the planned outcomes and the actual results in the classroom.

When this technique was adopted and applied over a period (in each case two full school terms), the result seemed to be that teachers involved in the experiment became more adept at reflection and more capable of articulating their own strengths and weaknesses from lesson to lesson, as well as developing the ability to judge their performance and to mark the most significant changes in their developing abilities to implement the new curriculum and integrate new curricular materials into their lessons.

In terms of the sustainability of the projects and developments for the future, the sharing of this technique in focus groups and workshops seemed to suggest that it would be of use to individual schools and clusters of schools in peer observations and evaluations, as well as by others within a particular school environment, to serve as a simple tool for providing informal feedback on the developing skills of individual teachers, and in mapping the most significant changes and milestones in their professional development. To this end, a personal portfolio was established for each teacher participating in the experiment, where the photographs and (their own) comments and reflections could be stored and referred to over a given period.

Conclusion

Feedback from all those participating was generally favourable, and Web-based analysis through automated polling (through an e-Languages project interface) suggested that the experiment was, in each case, a worthwhile point of departure for further experimentation of this sort. No specific statistics were gathered to support the informal polls, however, and further research into techniques of reflection and self-analysis of performance are clearly indicated by this pilot.

12

Teachers' voices on the washback effect of the high-stakes national examination in Indonesia

Afrianto

Abstract

The National Examination (UN) policy has been a public debate among educational practitioners and policy makers in Indonesia of recent times. Those who oppose UN argue that there is an inherent 'injustice' in applying one examination within a subject area across the whole of the country, the results of which will ultimately impact on the students' future life. The government, on the other hand, is keen to pursue UN as a means of evaluating the results of teaching and learning processes across the country. This qualitative study tried to investigate teachers' voices in terms of the perceived washback of the UN. The analysis of in-depth interviews done with six English teachers as participants of the study shows that the UN has led teachers to teach to the test, made the teachers as well as students feel stressed and under pressure; pushed the students to engage in cheating; and narrowed the curriculum.

Introduction

National Examination (UN) policy for standardised testing for secondary (lately also for primary) school students in Indonesia has triggered a hot national debate since the beginning of the 2003/2004 academic year. Those who oppose it argue that this policy is considered to be 'injustice', to be used as the basis for making a very important decision about students' future lives. This is due to the fact that there is still a big discrepancy in quality among schools across the regions in Indonesia. The government, on the other hand, says that the UN is important as the government needs it as a benchmark to evaluate the success of teaching and learning process at national level.

It is necessary to conduct a thorough study on what sort of impacts the UN has had on teaching and learning processes at school. The later findings are expected to be used by the related parties in finding out a way out of this issue by forming a relatively acceptable format of UN in Indonesia.

Some Important Features of Current National Examination

As a national standardised test, the UN is addressed to all high school students who sit in the third year (the new term used in the latest curriculum is "grade twelve" for senior high school or 'grade nine' for junior high school) of their schooling period. Initially there were only three subjects tested in this UN (Bahasa Indonesia, English, and Economic for Social Science Students), but starting from the 2007/2008 academic year, the government decided to include three more subjects. The new ones are Maths, Sociology, and Geography for Social Science students, or Biology, Chemistry and Physics for Natural Science students (Depdiknas, 2007).

According to clause 2 of the Decree No. 34/2007 from the Ministry of National Education or *Permendiknas*, the main goal of the UN is to measure and assess the students' knowledge and competence in particular subjects they have learned. One of the important characteristics of UN is that the government employs the minimum threshold (popular with passing grade) for the candidates to achieve in order to pass the examination. The minimum threshold is increased year by year, from 3.01 in 2003 to 5.50 in 2010.

Again, the candidates must achieve the minimum threshold in order to pass the test. Otherwise, they are considered 'failed'. Consequently, they have to repeat all subjects in the following academic year. In other words, failure to achieve the minimum threshold in UN will automatically result in failure to graduate from high school, regardless of the student's overall performance during their school years.

This UN is powerful in determining students' future lives as it functions as a 'gatekeeper', which will allow or not allow the candidates to pursue their studies further. That is why this national standardised test can be classified as what McNamara has called a *"high stakes test"* (2000, p. 48).

Washback Effect

Washback refers to the influence of testing on teaching and learning (Bachman & Palmer, 1996; Cheng, 1997; Gates, 1995 in Brown, 2002; McNamara, 2000). For Messick (1996, cited in Brown, 2002 para. 9), washback is, "the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning".

Research Methodology

This study employed a qualitative research approach. This research was carried out with six experienced Indonesian English teachers aged 25 to 40 years old who are teaching in the third grade (or grade twelve according to the latest curriculum) of senior high schools in Tanah Datar District, West Sumatra Indonesia. The participants were chosen by criterion purposeful sampling (Patton, 1990). Grade twelve teachers, rather than teachers from other grades, were selected for the study on the grounds that, because they have to help their students prepare for the UN to be taken at the end of school year, they would be more concerned with and have a better knowledge of the philosophy of the test than teachers teaching the other grades. Therefore, their input was considered potentially valid for this study. Considering the ethical issue, anonymity is used for all participants' names in this study.

Results and Discussion

There are three dimensions to the impact explored during the interviews; they are the impact of the test on instructions and curriculum, the impact on teachers, and the impacts on students. The data analysed from transcribed interviews show us some identified negative and positive effects of the UN as perceived by the teachers. The effects are as follows:

1. Teaching to the Test

From the data collected, it is apparent that the high stakes of UN has led teachers to teach to the test. This activity implies doing something in class that may not be compatible with the teacher's own values and goals or with the values or goals stated in the curriculum. Most teaching activities focus on familiarising the students with the features of the test as well as introducing test taking strategies to enable them to answer the questions well. The teaching to the test phenomenon has also made teachers neglect other subjects which are not tested in the UN. All participants in this study reported that they conducted extra classes where most of the time they employed activities like familiarising the students with the test format, discussing the questions, discussing strategies to answer the questions more easily and more quickly as well as conducting some trial tests prior to the real examination.

It is worthwhile bearing in mind that the practice of teaching to the test could bring about some problems. It could result in some unwanted consequences within the nature of teaching and learning. Popham (2001, cited in Volante 2004) has argued that "item-teaching, instruction around items either found on a test or a set of look-alike items, is reprehensible since it erodes the inferences we can make about students' scores." We cannot simply judge a student's English proficiency, for example, merely based on his or her English score in the UN. A student who gets a high score after being exposed extensively to items of the English UN through items-teaching activities might have poor real English proficiency. On the other hand, it is possible for a certain student who has relatively good English competence to get a lower score, because the teacher does not employ items-teaching, and therefore the student is not familiar with the test mechanism.

In this context, Volante (2004, para. 8) further reminds us that research conducted by Shepard (2000) and Smith and Fey (2000) suggests that, "while students' scores will rise when teachers teach closely to a test, learning often does not change. In fact, the opposite may be true. That is, there are schools that have demonstrated improvements in student learning while their standardised test scores did not show significant gains."

This research finding implies that a high score obtained by students in a particular school might not accurately reflect that a school has a good teaching quality. It is possible that they get a good score, because 'teaching to the test' activities are used extensively prior to the test. Conversely, it is likely for the students who enrol in a school with a good English programme to get a lower score, because English teachers in this school focus on the nature of teaching as mandated in the English curriculum, instead of focusing on teaching to the test. Furthermore, the practice of teaching to the test in Indonesian classrooms has also undermined the predictive validity of the test results, as the results are likely not to give an authentic picture of the candidates' proficiency, and therefore could not be used as the basis to predict their academic achievement in the higher levels of education. Thus, "the predictive validity of a standardised test is compromised when teaching to the test techniques are employed" (Burger & Krueger, 2003 cited in Volante, 2004, para. 11).

It is worth nothing that when school graduates want to continue their study to higher levels of education in Indonesia, they are required to take another test for entering the universities (well-known as *The University Entrance Test/*SMPTN) as the authorities in universities are unwilling to take the UN score results into account. In other words, the students' scores from this UN are 'useless' to predict students' future life in education.

2. Narrowing the Curriculum

Another subsequent impact of teaching to the test activities as perceived by teachers is that the UN, in some ways, has narrowed the school curriculum (Yeh cited in Mitchel, 2006). This means that the teachers mainly focus on teaching the subjects tested in the national exam and tend to ignore other untested subjects. In current Indonesian classroom practice, time is often taken away from subjects like history, religious teaching, physical education, arts, and Information Technology. In other words, teachers provide more instructional time on commonly tested areas like Bahasa Indonesia, English and Mathematics.

Ignoring the untested subjects in the schools could undoubtedly lead teachers to narrow down the curriculum. For one thing it tends to communicate a misapprehension that the other subjects are not as important as other tested-subjects. Furthermore, a serious problem may appear if teachers as well as students think in such a way, since they may find in their real life later that the ignored subjects are, in fact, very important. In the English teaching context, a student may develop a narrow view of English learning. They might have been misled by the fact that the English test in the UN only addresses two macro skills (reading and listening), and therefore many teachers focus on teaching these two skills. It is possible that this focus would lead students to think that the other skills (speaking and writing) are not as important.

3. Willing to Engage in Cheating

The high stakes nature of the test has encouraged some students in Indonesia to engage in cheating during the examination. The cheating itself is not only triggered by the high stakes nature of the test; issues of unfairness in the passing grade policy have also contributed to the cheating phenomenon during the UN in Indonesia. One participant of this research confessed that her students were engaged in cheating, because they had to achieve the threshold in order to pass the test, otherwise they would need to repeat it in the following year.

Cheating cases in the UN have been identified in some other schools across the regions in Indonesia. Some cases appeared to the public when the UN was conducted in 2006/2007 academic year. It was reported, for example, that 72 of Dhuafa Vocational High School students in Padang West Sumatera walked out from the test rooms as a protest to the exam committee. They could have perceived the committees as doing nothing when other students were allegedly cheating in the examination (Bachyul, 2007).

Another case was in Medan City, North Sumatra. Some teachers in this city quit from being test invigilators and then gathered to report alleged systematic cheating all over the Medan region. This group of teachers attracted nationwide attention when they presented evidence of rampant cheating during the examination. They reported that the cheating itself had been systematically organised by some principals and teachers long before the test day (Gunawan, 2007).

It is believed that these cheating cases are closely related to the issues of unfairness within the passing grade policy. As the threshold is considered too high for their students to achieve, some school principals might try to find out a 'shortcut' to pass the test. They do not want to see their students fail in the test, because if many students fail, as school principals, they are going to be the first persons to be blamed by parents and society.

4. Stressed and under pressure

The high stakes nature of the test has made teachers feel stressed and under pressure in conducting teaching activities prior to the test day. This stress is also triggered by the fact that school principals and parents have high expectations which are transferred onto the teacher in order to help students pass the test. Consequently, many participants in this study reported that they were feeling insecure and worried if their students did not pass the test. They are afraid of being branded as unqualified teachers if many students fail in the examination.

This finding is consistent with those from other studies about the impact of high-stakes testing on teachers, such as a study by Wright in 2002 to examine the effects of the SAT-9 on a large inner-city elementary school in Southern California. Apart from narrowing the curriculum, the study reveals that standardised testing resulted in harmful effects on both teachers and students. One of the effects on teachers is that, "teachers are stressed and overwhelmed by all the curricular changes and pressure to teach to the test and raise scores". (p.28)

It is obviously not good for the teaching process if teachers feel under pressure for the reasons outlined above. This insecurity may lead them to a situation where they cannot enjoy their profession. The worst thing is that this unwanted situation will eventually affect educational quality in Indonesia. It is certainly a paradoxical situation, as the existence of the UN itself was initially intended to improve the quality of national education in Indonesia.

5. Positive Impact of the UN

Apart from negative impacts outlined above, this study has also revealed that the test has brought some positive effects to the teaching and learning process. The most salient one is that it has made most teachers as well as students invest more efforts into the process of teaching and learning. Most participants reported that the test has made them become more motivated to teach better, more creative in finding out enhanced teaching strategies, and more efficient in managing the teaching time allocation. At the same time, teachers reported that most of their students are also motivated to study harder and to use their time to study wisely.

The fact that this test has improved teachers,' as well as students', motivation to teach and study harder is used by the government to maintain the UN policy. Apart from using the UN as a means of controlling national education quality, the government also argues that this is one of the effective ways to make sure that teachers do their best, and students put substantial efforts into learning (Kompas, 2005).

Conclusion

This study reveals that the high stakes of the UN has affected instruction in a generally negative way. The effects are that the test leads teachers to teach to the test; it tends to narrow the curriculum; makes teachers stressed; it encourages students to engage in cheating, and encourages teachers to have so called 'score oriented teaching'. However, the study has also found that the test has encouraged teachers as well as students to teach and study harder in order to pass the test.

References

Alderson, J. C., & Wall, D. (1993). Does Washback Exist? ***Applied Linguistics, 14(2), pp. 115-129.***

Bachman, L. F., & Palmer, A. S. (1996). ***Language testing in practice : designing and developing useful language tests.*** England: Oxford University Press.

Bachyul, S. (2007). Walked out students not allowed to sit repeat exams. ***The Jakarta Post (May 5).***

Brown, J. D. (2002). Extraneous variables and the washback effect [Electronic Version]. Shiken: ***JALT Testing & Evaluation SIG Newsletter, 6, pp. 12-15.*** Retrieved 27/01/2007 from http://jalt.org/test/bro_14.htm.

Cheng, L. (1997). How Does Washback Influence Teaching? Implications for Hong Kong. ***Language and Education, 11(1), pp. 38-54.***

DeCesare, D. (2002). How High Are the Stakes in High-Stakes Testing? ***Principal - The Standardised Curriculum 81(3).***

Depdiknas. (2007). ***Peraturan Mentri Pendidikan Nasional No.34 Tahun.*** Retrieved. from http://www.depdiknas.go.id.

Gunawan, T. S. (2007). National Exam Encourages Teaching. ***The Jakarta Post (May 19).***

Jacob, B. A., & Levitt, S. D. (2003). Catching Cheating Teachers: The Results of an Unusual Experiment in Implementing Theory. ***Brookings-Wharton Papers on Urban Affairs, pp. 185-220.***

Kompas. (2005, January 31). ***Ujian Nasional Jalan Terus. Kompas.***

McNamara, T. (2000). ***Language Testing.*** New York: Oxford University Press.

Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment ***Educational Researcher,*** 18(2), pp. 5-11.

Mitchel, R. (2006). ***Research Review: Effects of High-Stakes Testing on Instruction*** [Electronic Version]. Retrieved 24/01/2007 from http://www.centerforpubliceducation.org/site/c.kjJXJ5MPIwE/b.1536671/k.9B6A/Research_review_Effects_of_highstakes_testing_on_instruction.htm.

Napitupulu, E. L. (2007, May 04). ***Kecurangan UN Justru Terjadi di Ruang Kelas. Kompas Cyber Media.***

Schrag, F. K. (2004). High Stakes Testing and Distributive Justice. ***Theory and Research in Education, 2(3), pp. 255-262.***

Shohamy, E., Donitsa-Schmidt, S., & Ferman. (1996). Test impact revisited: Washback effect over time. ***Language Testing, 13, pp. 298-317.***

Sloane, F. C., & Kelly, A. E. (2003). Issues in High-Stakes Testing Programs. ***Theory Into Practice, 42(1).***

Volante, L. (2004). Teaching To the Test: What Every Educator and Policy-maker Should Know. ***Canadian Journal of Educational Administration and Policy (35).***

13

Assessment for motivation: Incentives for teacher professional development

Vu Mai Trang

Abstract

That recent research increasingly recognises the role of teachers as the most important factor in student achievement further emphasises the need for teachers to develop their professional learning as a lifelong skill, and for administrators to include this in teacher evaluation as a quality assurance tool. However as professional development (PD) involves independent learning, which is highly dependent on teacher motivations, assessing this seemingly non-assessable work could have a counter effect if set on a rigidly compulsory and judgmental basis. The paper introduces the role of assessment, in the light of motivation, in supporting teacher PD and is demonstrated with findings from a framework recently conducted in a university in Viet Nam. It explains why this 'assessment for motivation' was able to boost teacher motivation in their learning and promote sustainable PD. It is hoped to offer some helpful insights into assessment and its link with educational improvement that can be used in similar pedagogical situations.

For me, teaching is not tranquil. Every time I enter my classroom my heart beats as if it was my first time. I see a multi-layered mixture of fulfillment, passion, success, fear, satisfaction, loneliness, ambiguity, frustration, and above all, the uncertainty about if I can live up to the expectations from others.

Paulo Freire in his acclaimed Pedagogy of Hope (p.69), states his view on what it is like to be a teacher:

> *"Teachers who fail to take their teaching practice seriously, who therefore do not study, so that they teach poorly, or who teach something they know poorly […]. Thus, they disqualify themselves as teachers."*

Although his view might be extreme to some, it is important that before teaching somebody to learn, a teacher should be a learner herself, who constantly embarks on new journey of discovering and learning.

Teacher learning is essential because this has a wide impact on others. Research has proved the interrelationships between teacher learning and student performance (Smith & Gillespie, 2007; Desimone, 2009). This further emphasises the need for teachers to develop their professional learning as a lifelong skill, and school administrators to include this skill in teacher assessment as a tool for quality assurance.

The dilemma

However, as professional development (PD) greatly involves independent learning, which depends to a great extent on teacher motivations, assessing this seemingly non-assessable work could have a counter effect if it is set on a rigidly compulsory and judgmental basis. In their study Schieb & Karabenick (2011) list some reasons for teacher low participation in PD including teachers' resistance to educational reform and teacher level of dedication and motivation.

For Vietnamese teachers, let's look at the following vignettes:

> *"I've been working as a teacher for 25 years and I almost feel burn-out at work. I think I am one of the luckiest teachers who have attended the *…+ course."*
>
> *"I've been teaching for quite a long time, but I'm not satisfied with my teaching now. I believe that I must improve myself at once, otherwise, teaching will no longer an enjoyment… I want to make my teaching more interesting and more beneficial to my students."*
>
> *"Most Vietnamese teachers are busy with not only teaching but other family and social issues. They have to work extra time to support their life."*
>
> *"Some teachers do not really care.. nobody can fire them even if they do nothing to improve their knowledge and skills."*

Two dilemmas over PD can be interpreted from these accounts: one from the teacher's perspective and one from the administrator's perspective. As the person responsible for PD and teacher development programmes for a faculty at Vietnam National University (hereafter "the Faculty"), I face the question of helping to make PD here more effective.

The purpose of my enquiry

In order to be better informed, I conducted a survey in May 2010 on the current situation of teachers participating in PD at the Faculty.

The status quo investigation

The Faculty has an in-house conference annually, and every teacher is expected to be the author/co-author of a paper. However, teachers complain sometimes they just write up the paper by filling pages with words, and the papers are often left untouched after the conference ends. There are two or three seminars for teachers in a year, but it seems the topics are not always relevant to their needs and interests. This is understandable because the seminars are organised for big audiences from all content areas in the University.

The survey

Thirty-six teachers were invited to take the survey. These teachers are both new and seasoned teachers, having 1-32 years teaching experience. From the survey it was found out that not every teacher at the Faculty has concrete ideas of what form PD can take and how to do it. When given a list of activities (all of which are in fact for PD) and asked which activities they think are for teacher learning, only 1 in every 3 teachers (33%) correctly ticked all the options. The rest do not see activities such as team teaching, journal writing, and reading as learning opportunities for teachers.

All of them think PD is just formal activities such as doing workshops, submitting papers, giving presentations at conferences, and so on.

When asked to tell what they have done, or have been doing for PD, most of them ticked formal PD (doing MA, PhD). The other most selected activities in the list are discussions with colleagues and students, and class observations. Few have been taking ongoing learning activities (e.g. short courses), self-learning and reflection activities, and writing research papers. They specified the obstacles to PD: “too difficult”; “don’t have enough time”; “haven’t got anything that’s really of interest”; “I’m not yet confident enough”.

With these findings, it can be concluded that even if there are some PD forms at the Faculty (research, seminars), the teachers are still reluctant to participate. This may be because they do not see the utility of these activities, as well as the relevance with their own concerns and interests. The fact that there is a gap between research and practice, which in fact may be quite common elsewhere, also contributes to this low motivation. Another reason that leads to low participation in PD is teachers do not have sufficient knowledge of what PD is and how to do it. They only associate PD with formal activities like doing research and delivering papers, but not informal activities including reflection, which limits their own access to learning opportunities. Besides, it seems they do not have a strong feeling about the need to develop themselves professionally. For those who want to do so, that there has not been an assessment/recognition mechanism and this may discourage them. The fact that they stick mostly with PD activities like discussion and observation, and ask for refreshment training courses, shows that although they find doing research quite beyond their reach, they want to be empowered to do so.

Issues to be addressed

With these findings, it seems that the issue that needs addressing is to set up a PD assessment framework, hereafter called “the Framework”, that:

- Demystifies “Professional Development”, including classroom research
- Motivates teachers intrinsically and extrinsically in taking on PD
- Promotes PD and learning opportunities for teachers through evidence-based practice, and
- Serves as a tool for teacher assessment and quality assurance.

Approaches and Principles

In setting up such a framework, there are three approaches and principles I follow:

1. Connect Extrinsic & Intrinsic motivations with PD with the introduction of Teacher standards and Teacher values.
2. Use Positive washback effect/Backward design in the making of the Framework.
3. Introduce various Professional Development notions and concepts.

1. Connect Extrinsic & Intrinsic motivations with PD
To motivate teachers extrinsically, I suggested doing PD as part of evaluative criteria for teacher assessment as one of the trinity of tenure criteria: scholarship, teaching, service. Moreover, doing PD can be embedded in foreign language teacher standards that are being proposed in Viet Nam. Figure 1 summarises these standards and shows a strong connection with learning through PD (adapted from Bransford, Darling-Hammond & LePage (2005) & Ball & Cohen (1999) in Dudzik, 2010).

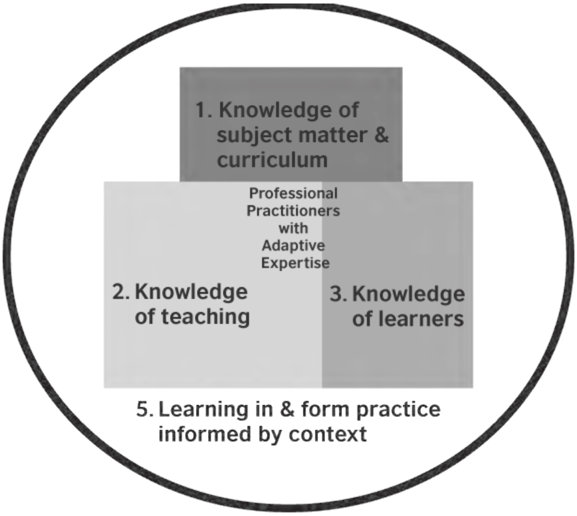

1. Knowledge of subject matter & curriculum
Professional Practitioners with Adaptive Expertise
2. Knowledge of teaching
3. Knowledge of learners
5. Learning in & form practice informed by context

Figure 1 - Teacher Standards.

To motivate teachers intrinsically, it is necessary to connect PD with teacher values. Teachers need to be aware they are professionals, not lay, amateur, technicians, nor academics. Ur (2002) defines the teacher as a bringer-about of real-world change, which requires lifelong learning and development, and this is not always easy. Brookfield (2006) put this in the three R's of skillful teaching: respect, research, and responsiveness.

2. Use Positive Washback effect/Backward design in the making of the Framework
In designing the Framework, I used washback effect as one of the principles. To achieve beneficial washback is to test the abilities whose development you want to encourage (Hughes, 1989). "Backward" design was also used. Similar to washback, "backward" design, (cf. Wiggins & McTighe, 2005), involves the practice of looking at the outcomes in order to design curriculum units, performance assessments, and classroom instruction. And before any assessment is made, assessment criteria should be clearly communicated to teachers, following Hughes's (1989) principles for beneficial backwash effect.

3. Introduce various Professional Development notions and concepts
This relates closely with demystifying PD notions and concepts, making them more teacher-friendly and less frightening. In the Framework a number of PD activities are introduced, and they are both traditional ones (learning seen as the transmission of knowledge) and reflective ones. Both formal PD practice (conferences, workshops, courses, academic study), and informal learning

opportunities (journal writing, critical incidence analysis, team teaching, etc.) are included, based on models by Richards and Farrell (2005) as I find them quite accessible to teachers who begin to take on PD. Noticeably, both the informal and formal PD may lead to action research at the end. Also, PD needs to be understood as a process, “ongoing, coherent, and continuous, rather than unrelated and episodic” (Myers & Clark, 2002, p.50). It is not the amount of PD that matters, but the process of development and the quality of that process that is essential for changing practice and professional competence (Murray & Christison, 2011).

The Framework

The Framework comes in the form of a grid for teachers to fill in. Teachers receive this at the beginning of the school year and submit the report form at the end of the year, certified by their head of division, and may supply further evidence if requested. Results from the report are to be used for purposes such as contract granting and tenure possibilities.

It should be noted that this form itself does not make up the Framework. We organised INSET, working groups, mini conferences, thematic seminars and workshops, and an E-journal. In other words, teachers are not only provided with a form to tick; they are provided with opportunities so that they have something to record (see Figure 2).

Professional Development Activities from … to … (please tick and give additional information. You are not expected to do ALL of these)		
Activities	Yes	Yes Please describe (e.g. title, date, venue, people involved)
Courses attended/am attending		
Materials development		
Publications (books, articles)		
Research papers		
Ongoing research		
Conference presentations		
Conferences/Workshops/ Seminars attended		
INSET (at least 3)		
Research Projects		
Committee/Working group		
New teacher coaching		
Class observations		
Reflections/Teaching journals		
Books/Articles read		
Self-study activities		
Other Professional Development Activities		

Figure 2 – The Framework Form.

Why the Framework may assist teacher development

- As it suggests possibilities for our development, the Framework encourages learning opportunities, makes PD less frightening and more teacher-friendly. While it was found out that most teachers at the Faculty associate PD with writing up a research paper, the Framework suggests other alternatives, covering both formal and informal PD. This is especially important for the young faculty who are starting their careers. For those who are accumulating scholarship credits and credibility the Framework provides them a possibility of being acknowledged, as well as esteem and self-actualisation as described in Maslow's hierarchy of needs and motivation.
- The Framework covers almost all aspects of contemporary views on PD, including content focus, active learning, collective participation, learning through experience, and learning from reflection (Desimone, 2009).
- As a tool to trace our growth trail, it helps restore our vigour and enthusiasm, reminds us we are professionals with qualities and responsibilities different from others', as Ur (2002) mentions.
- It creates a sense of belonging to a community. With this Framework teachers will feel they are in a secure learning environment, by knowing others are also doing what they are doing, and by doing things with others. Burnout, which leads to leaving the profession, is lessened if teachers have peer support – "peers can provide help, comfort, insight, comparison, rewards, humor and escape" (Barduhn, 2002).
- The Framework encourages teachers to take reflection even to a more meta-cognitive level, when they detach from themselves, looking at what they have done, are doing, and will do in terms of self-development. This "out of body" experience is essential because PD is of long-term basis and hence in this process meta-cognitive strategies are often used.

Trialing: Case studies

The form was sent to four teachers at the beginning of Spring semester (Jan 2011) to fill in. It was sent again to the same four teachers at the end of the semester. Comparing the two reports by each teacher, it was found out that their PD activities were much more abundant and various (seminars/workshops attended, books/articles read). Further communication with them shows that they are keen on this framework, and feel more motivated with PD. The young male teacher when submitting the first report told me that he was ashamed as he was doing little on PD, and he said his form was almost empty. The second report by him shows a much more confident teacher, with a lot more PD activities and practices. In addition, the teachers said it would help if they received more guidance on the activities in the form. In response to these concerns, a workshop specifically on PD models and demonstrations was conducted in May 2011 at the Faculty as an effort to bring teachers, especially the junior ones, tools for their own PD plans.

Challenges

There is much debate on how PD should be measured and assessed. It is indeed questionable if PD is seen as a quantifiable list of activities. Again, as teacher professionals focus on real-world change, PD needs to be sustained and intensive and focused on the actual classroom. In other words, their knowledge store is enriched, but they may not act. It is essential to apply such knowledge to one's own context.

Smith (2010) says we should couple PD and professional learning activities with practices and policies that support what teachers do and how students learn. Nir & Bogler (2008) also find that PD programmes are most beneficial when teachers maintain input and control in the PD process and are linked to the participants' teaching culture, curricula, and classrooms.

In improving the Framework, there are some major challenges that need to be addressed, including:

- More evidence on change in teachers' PD behaviour
- How to better bridge the gap between research and practice
- How to maintain teachers' motivations once they have been created
- How to better measure the success.

Moving forward

As the Framework is indeed a mechanism of formative assessment, building it as, for example, a standard-referenced system may resolve the conflict between formative and summative assessment when it comes to teacher evaluation and quality control.

References

Barduhn, S. (2002). 'Why develop? It's easier not to'. ***Continuing Professional Development.*** Ed. Julian Edge. IATEFL, 2002, cited in Murray, D. E. & Christison, M. (2011).

Brookfield, S. D. (2006). T***he Skillful Teacher. On Techniques, Trust, and Responsiveness in the Classroom.*** Second Edition. San Francisco, CA: Jossey-Bass.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualisation and measurement. ***Educational Researcher, 38 (3), pp. 181-199.***

Dudzik, D. (2010). Vietnam Foreign Language Teacher Qualifications Framework Proposal. Vietnam Ministry of Education and Training.

Freire, P. (1992). ***Pedagogy of Hope: reliving Pedagogy of the oppressed.*** The 2004 Edition. London: Continuum.

Hughes, A. (1989). ***Testing for Language Teachers.*** Cambridge: Cambridge University Press.

Murray, D. E. & Christison M. (2011). ***What English Language Teachers Need to Know.*** New York, NY: Routledge.

Myers, M. & Clark, S. (2002) 'CPD, lifelong learning and going meta'. ***Continuing Professional Development***. Julian Edge (Ed.). IATEFL, 2002, cited in Murray, D. E. & Christison, M. (2011).

Nir, A. E. & Bogler, R. (2008). The antecedents of teacher satisfaction with professional development programs. ***Teaching and Teacher Education, 24(2), pp. 377-386.***

Richards, J. & Farrell, T. (2005). ***Professional Development for Language Teachers.*** Cambridge: Cambridge University Press.

Schieb, L. J. & Karabenick, S. A. (2011). Teacher Motivation and Professional Development: Review and Knowledge Database. University of Michigan. US National Science Foundation funded project. Retrieved from ***mspmap.org/wp-content/uploads/2011/.../TeachMotivPD_TechReport.pdf*** on 17 June 2011.

Smith, C. (2010). The great dilemma of improving teacher quality in adult learning and literacy. ***Adult Basic Education and Literacy Journal 4 (2), pp. 67-74.***

Smith, C. & Gillespie, M. (2007). Research on professional development and teacher change: Implications for adult basic education. In J. Comings, B. Garner, & C. Smith (Eds.) ***Review of adult learning and literacy Vol. 7, pp. 205 – 244.*** Mahwah, NJ: Lawrence Erlbaum.

Ur, P. (2002). The English Teacher as Professional. In J. Richards & W. Renandya (Eds.). ***Methodology in Language Teaching: An Anthology of Current Practice, pp. 388-392.*** Cambridge: Cambridge University Press.

Wiggins, G. & McTighe, J. (2005). ***Understanding by Design.*** Expanded 2nd Ed. USA: Association for Supervision and Curriculum Development.

14

Developing skills in communicative test writing - a China case study

Keith O'Hare

Abstract

This paper looks at recent changes in English teaching in China, and the impact the existing assessment system is having on classroom teaching in schools across the country, and then against this backdrop, it introduces a case study of a project run by the British Council in China that focuses on training for test developers. The paper explains why this project came into being, considers the successes and challenges of the project so far, and finally shares learning points that may benefit other organizations running similar projects.

Recent changes to English teaching in China

In the 1980s, China initiated its opening up policy and very quickly realised it needed to invest more in its education system and nurturing talent. Many people were sent abroad in order to learn about the latest developments in science and technology in other countries and then bring that knowledge back to China to help the country progress. Of course, the use of English played a big part in that. However, many people leaving schools knew about English grammar but few could actually use English to communicate. If China was to operate in a global world, to learn from other countries and to do business with other countries, then the way English was taught in schools needed to change.

National curriculum reform for English (2001)

So in 2001, the National Curriculum for English underwent a major reform. The reform aimed to create learners who could communicate in English, by creating teachers who could use communicative methodology and by having tests that assessed communicative teaching. Before 2001, the main component of the National Curriculum was language knowledge, i.e. knowledge of grammar and vocabulary. The new curriculum added four new pillars to the curriculum: language skills, affect and attitude, learning strategies, and cultural awareness. In addition, an underlying principle on assessment was added that stated that the guidelines for the national curriculum should emphasise formative evaluations.

All of the above were very ambitious aspirations, and led to a huge investment by the government in training for English teachers. Generally speaking, the approach was to train the best teachers, who were then expected to mentor, guide and give demonstration lessons of best practice to other teachers. From 2001 to 2011 some progress was made. Teaching in some schools of the bigger cities began to move away from a grammar translation approach to a more communicative approach. Text books were changed and those approved by the Ministry of Education included task-based learning activities, learning strategies, cultural awareness activities, as well as communicative tasks.

However by 2011, the majority of English teaching in schools had still not moved towards a learner-centred, communicative approach. Probably the biggest reason for this was that the assessment system did not encourage communicative teaching. Whilst the reform of 2001 emphasised formative evaluations, in fact this did not really happen.

Current assessment systems in China

Today in China a lot of teaching is controlled and guided by the assessment system which consists of three main, high-stakes, summative examinations. They are the examinations at the end of primary school (taken at around 11/12 years old), at the end of lower secondary school (taken at around 14 years old; known as the *"Zhong Kao"*), and at the end of upper secondary school (taken at around 17 years old, known as the *"Gao Kao"*). Whilst the examination at the end of primary school is not a national one and does not formally dictate which secondary school a child goes to, the reality is that a child's performance in that examination is often taken into account by secondary schools.

The two big high-stakes examinations are the *Zhong Kao* and *Gao Kao*; the first decides which upper secondary school you will get into and the second determines which university you can get into.

These two high-stakes examinations, by and large, remain discrete-item focused, with a strong reliance on multiple choice items, and have items that usually give insufficient context. The result is a negative washback on teaching in classrooms in secondary schools across the country. It is no exaggeration to say that most of the class time in upper secondary school is not spent on learning English, but rather on practising discrete-item tests in order to prepare for the final end of school examination. The problem is not just limited to these two examinations. Since their influence is so powerful, many school tests, such as monthly tests or mid-term tests, also follow exactly the same style as the high-stakes examinations. Consequently, many students may not be interested in English classes, have no motivation and may not become effective communicators of English. Without a doubt, the backwash of these examinations and tests is one of the major barriers to communicative teaching in China.

So why have these high-stake examinations not changed? Interestingly, there have been some changes in these examinations and they are starting to introduce more communicative items, however, the changes are limited; speaking for example, is not tested. The main reason seems to be the huge numbers of students that need to be tested. With a population of around 1.3 billion, there are an estimated 300 million learners of English, the majority of which are in the education system, between kindergarten and university. Having one summative examination based on objective test items, it is suggested, is the fairest and most effective way of assessing so many people.

Challenges for teachers and test designers

In 2009, the British Council undertook research into the needs of teachers in China, and the following information was gathered about teachers and teacher researchers. In China, a teacher researcher is someone who was a good teacher and has been given a new role to mentor new teachers, observe classes and give feedback, organise training for teachers, and write examinations for teachers.

Teacher researchers told us that they felt high-stakes examinations have had little change and have been “directing” classroom teaching since 2001. They also said they would like more training on how to design examinations and learn more assessment theory. Interestingly, most teacher researchers do not get formal training on test design. Many will learn it on the job, often drawing on models they have received in the past as a student or as a teacher.

They also said that teachers face many challenges and problems regarding assessment. For example, they may not be sure what a test item really tests, and they may not be sure how to interpret examination results. Also, they mainly use tests to evaluate students; there is very little assessment for learning; so students are given a mark but no guidance on how to improve their learning. Finally, many teachers, being busy, take tests from magazines and newspapers. There is a plethora of newspapers available for teachers and students to help them learn English. Many contain ready-made tests that teachers can photocopy and use. However, the quality of these tests, including reliability and validity, are often deemed as dubious. The biggest problem for teachers is that they lack the skills to evaluate and judge the quality of these tests, so they may use them indiscriminately.

British Council Examination Design and Assessment Project

In that context, the British Council set up a project to give test designers and teachers the skills to create effective examinations that would have a positive washback on communicative classroom teaching. In turn, this would help students become better communicative users of English, so meeting the requirements of the National Curriculum.

The project is run with local Chinese education authorities in a variety of cities across the country. It is a 5-6 month skills development course involving face to face training and online practice integrated into the work context. It is aimed at teacher researchers of lower secondary schools.

There are two stages that have the same format but different content. The structure can be seen in Figure 1 below.

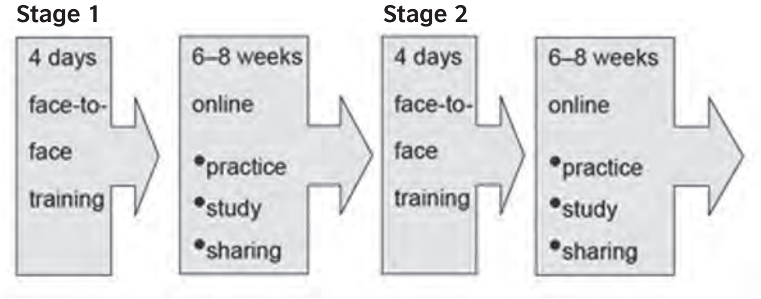


Figure 1 - Project Stages.

The content of the stage one face to face training includes theory behind good assessment tools, UK approaches to testing and assessment in secondary schools, evaluating test papers (UK and Chinese ones), writing test specifications, and writing test items. This is followed by online training, which was initially run on a public file-sharing site called Huddle. The online activities included a review of the face-to-face materials that were put up on the site, reading articles and watching video clips, test writing done in small groups, and sharing tests on the site so that other participants could see them and comment on them.

The content of the stage two face-to-face training included reviewing the items and examinations made online, then creating test items for the four skills: writing, reading, listening and speaking, and finally looking at testing integrated skills. The stage two online training was moved onto a Moodle platform and included a review of face-to-face materials, carrying out a series of tasks such as writing tests and then trialling them out in class and sharing the results in the forums.

Learning Points from the project; working with partners

When working with partners we realised that we had to change the name of the project. The Chinese word for “examinations” was proving to be too sensitive. Examinations, or “Kao Shi” in Chinese, is the same word used for the national high-stake examinations, and so many partners assumed we wanted to make changes to those examinations, which was not the case.

Also, we originally wanted to work with the top level teacher researchers and have an impact on those people who were making provincial level examinations; however, the government control of those teacher researchers at provincial level was too strong. In fact, it was often impossible to find out who these people were. When they are used to write the high-stake tests, their identity is kept confidential for security reasons. So we adapted the project in the early stages to take a grassroots approach and by giving training to lower level teacher researchers. By show casing the impact we could have on test designers of low-stake tests, such as monthly tests, we could work towards engaging with educational partners at a higher level in the future.

Learning Points from the project - working with participants

Some participants found the course very challenging. Since most teacher training has very little training on assessment, we were introducing ideas that were difficult concepts for many. What is more, although they may have known some terms in Chinese, they were not always familiar with the English word. In order to support participants, we offered some pre-course reading tasks to get them familiar with the concepts, and we created a terminology list in both English and Chinese that was further developed during the training.

The online study created several challenges. In the beginning we used a Huddle platform, but this proved to be messy. Participants could upload their own files, and despite specified sections for certain files, it often got out of control. There was no sense of continuity and participants expected a sense of progression that the site did not give. Our answer was to change to a Moodle site and set up an online course with a sense of progression. Participants could still write items and these could be shared in the forums or on the wikis.

Another challenge was that most participants were not used to studying online and often lost touch with the course. To manage this, we had the participants work in teams and each team had a leader responsible for organising the work of the others. Also, with the introduction of the Moodle platform we had an e-moderator who could engage with the participants, encourage them and monitor their work. Whilst this was a big step forward, the online study area still remains very challenging and some adaptations, such as setting up the online community before the face to face training begins, are presently being trialled.

Finally, at times the participants became disillusioned with the usefulness of the training:

> *"after all when I get back to work, the influence of the "Gao Kao" will take over again, washback, won't it?"*

Because of this, for every training course, we had a local testing expert who facilitated the training and acted as a kind of change agent, keeping the participants motivated and willing to try to make changes to their testing.

The project so far has run with five local partners and through many adaptations that have been made to meet the local context. Feedback has been very positive. We are in the process of measuring the impact of the training courses beyond what is learnt and felt in the training, to what changes are actually made in writing tests at work, to the delivery of tests and the washback of such tests on classroom teaching.

We realise that measuring such impact is challenging and, like the National Curriculum Reform of 2001, this is a hugely ambitious project, but we are confident that working on making assessment more communicative, is the one area that will probably have the biggest impact on communicative teaching.

SKILLS ASSESSMENT

15

The elusive skill: How can we test L2 listening validly?

John Field

Abstract

There has been much recent discussion of cognitive validity: i.e. whether the mental processes employed by a test taker during a test resemble those that the same person would employ in a real-world context. This article describes the processes that contribute to listening; and suggests how they might form a framework for more valid second-language tests of the skill. It may prove hard to revise major international tests of listening to represent the skill accurately and proportionately. But an opportunity exists at local level for testing that is more focused and that sheds useful light on where learners' problems lie.

1. Introduction

1.1 The mind of the test taker

This article reflects the growing interest in the test taker as an important factor in test design and validation. There has especially been a focus on cognitive operations - on the way in which the mind of a test taker operates when employing one of the four language skills (reading, writing, listening and speaking) under test conditions.

Recent models of language performance (see O'Sullivan in this volume) represent the cognitive demands upon a language user as combining knowledge of the language with the *mental processes* that enable a particular skill to be carried out. An important factor linking the two is the level of *expertise* that the language user has developed. It is this which enables the user: a) to draw efficiently upon knowledge of the language stored in the mind and b) to engage appropriate processes. In effect, it is another type of knowledge – knowledge of how to use a language to communicate rather than knowledge that a language has X or Y or Z in its grammar.

An expert is somebody who can use a skill *automatically* - in a way that is rapid and that does not demand a great deal of forethought. Just as an expert driver does not have to think about the process of changing gears, so an expert speaker constructs and produces a sentence without having to pause to think about the words or grammar being used. So a good test of one of the language skills assesses how far the expertise of the test taker has developed. The role of the test taker is especially important in high-stakes language tests that are used *predictively*: to show that an individual is capable of performing well in a particular job, class or academic setting. There is a responsibility on the test designer to ensure that the test produces behaviour in the test taker that is representative of the behaviour that happens in a real-world context. Clearly we cannot reproduce the circumstances of a real language encounter in the artificial environment of a test. But we need to find out if the *mental processes* that a test elicits from a candidate resemble the processes that he/she would employ in non-test conditions. This is what is referred to as *cognitive validity* (Glaser, 1991).

1.2 The elusive skill

Here I consider cognitive validity in relation to the listening skill, which poses special problems for language testing. Listening and reading take place internally, which means that we have to teach and test them indirectly – by asking questions. But we should never forget that the results obtained tell us about the product of listening or reading, not the *process* (how the test taker arrived at the answer). Of two students who gave the right answer, one may have achieved it by understanding 94 out of 100 words; another by understanding only 15 words and making some good inferences. As a result, it is very difficult for a teacher or tester to make use of test scores to diagnose and deal with specific problems (see Field, 2008a: Chap. 2 for an extended discussion of this issue).

Listening and reading are often tested using similar methods. But they are very different in the demands that they make of the test taker, and test designers should beware of making easy comparisons. The input is very different. Written words are in a standard form thanks to spelling conventions, whereas words in connected speech are subject to great variation. While there are gaps between each word in a reading text, there are few in connected speech and a listener has to work out where one word ends and the next begins. The process of listening is also very different from that of reading. Listeners cannot look back to check as readers can – they have to carry forward in their minds a recall of the conversation so far. Moreover, a reader can speed up or slow down according to the difficulty of a test, but a listener cannot.

1.3 Construct validity in tests of listening

In addition, listening is the most complex skill to test. A listening test usually involves other skills – candidates might, for example, have to read the items. This raises issues of construct validity. To what extent are we testing listening and to what extent are we testing reading?

Test designers often attempt to ensure that their test is a sound one by creating a graph of the scores showing that they fit into a *normal distribution* – with a few students at the extremes achieving the lowest and highest scores and a large number in the middle of the range. This shows that a test discriminates well between students; but it cannot be taken as sure evidence that a test tests what it aims to test. Imagine a listening test with a very easy recording but some complicated multiple-choice options to be read. It might well be that the normal distribution reflects the range of reading abilities of the students rather than their range of listening abilities.

The message here is a simple one. We should not just rely on checking the effectiveness of a test by piloting after the test has been written. Weir (2005: 13) argues strongly that we need to know much more about exactly what the test aims to assess before we start writing it. This information should shape our test design and (if we are testing one of the language skills) the range of different processes that the test covers. In the case of a test of listening, we need to know how an expert listener behaves (what is the target behaviour that test takers are working towards?). We also need to know what test takers actually do in a listening test and how closely it resembles natural listening. This means that listening test designers need a detailed understanding of:

a) the speech signal that reaches a listener's ear, and the problems it might cause to an L2 listener.

b) the processes that an expert listener uses in normal circumstances and the way they might vary in the case of an L2 listener.

Where can they find this kind of information? Well, in the case of a) by studying the phonology and phonetics of the target language from a receptive point of view. In the case of b), psychologists of language have built up quite detailed accounts of all four skills, which draw upon concrete evidence obtained in research. So they do not need to use intuition to guess what the skill consists of (as in a more traditional sub-skills approach); this is very much an *evidence based* approach to testing.

The remainder of this article provides a general overview of listening theory from these two perspectives and considers its implications for testing.

2. The listening process

Widespread use of the blanket term 'comprehension' has encouraged teachers and testers to think of listening as a single function, confined to reporting a meaning consistent with what a speaker says. In point of fact, extracting meaning from a piece of speech entails not one but three distinct types of operation (Field, 2008a: Chap. 7):

- **Decoding:** matching incoming sounds to words in the listener's vocabulary;
- **Parsing:** imposing a grammatical pattern upon groups of words derived from decoding;
- **Meaning building:** adding to the bare meaning by relating it to context; building various points of information into an overall picture.

3. Decoding

Teachers and testers of L2 listening are often reluctant to recognise the important role played by decoding. This is partly because of the old belief that the ability to recognise individual words in a piece of speech is not important, since any problems arising can be resolved by drawing upon what is referred to rather vaguely as 'context'. This argument does not hold water. The term 'context' as used here generally refers to the listener's recall of the recording so far. Clearly, if the listener has decoding problems, they will also affect the accuracy of any information carried forward from earlier, and will provide an unreliable basis for resolving problems of word recognition.

A second reason for this attitude is historical. Early approaches to the testing of listening focused heavily on phoneme and word recognition, making use of dictation and even minimal pair discrimination. It is unsurprising that in due course testers reacted against this very limited coverage of the skill; but the pendulum seems to have swung too far in the other direction. This is unfortunate: the fact is that many failures of comprehension originate at the level of decoding and often carry knock-on effects. A listener who misinterprets the utterance *I went to assist* *her as I went to a sister*, would create a representation for the recording which involved a female relative or a nurse – and would then persist in listening for further references to this individual, even in the face of evidence that a mistake has occurred (Field, 2008b).

Because testers have, for good reasons, abandoned methods that focus solely on low-level units of language, they sometimes claim that they are not concerned with decoding skills and only target 'comprehension'. But this is not what happens in practice. Test items quite often require the candidate to report at word level and thus test accuracy of decoding; in the case of gap-filling tasks, almost all the items may be of this kind. So one cannot criticise listening tests for not featuring decoding at all; but it seems clear that they do not include it in a systematic way that balances it against other functions.

A further point concerns how much of a recording a candidate can be expected to decode accurately. Testers of listening often choose recordings on the grounds that the transcript shows them to be at or slightly above the language level of the candidate. *But the issue is not what the candidate knows but what he/she can recognise when it occurs in connected speech.* This carries two important implications. Testers of listening need to work from the recorded material, not just the transcript. And they need to be sensitive to how perceptible words and structures are in a recording. A clearly articulated word carrying sentence stress will provide a much easier target than one that is reduced in form, brief in duration and/or low in volume.

4. Parsing

The challenge posed by parsing also tends to be overlooked. Listening takes place in real time, with sounds reaching the listener's ear syllable by syllable. There is no way of looking back and checking as there is in reading – which means that a gradually increasing string of syllables has to be held in a listener's mind until it is possible to recognise a grammatical pattern in them (Field, 2008a: Chap. 11). As a grammatical unit builds up, listeners make certain assumptions based on probability. For example, hearing *The lawyer questioned…*, they expect to hear a direct object next *(probably the witness)*. If what comes instead is *by the judge*, they have to quickly abandon the grammatical pattern they provisionally allocated. As with the *assist her example*, this illustrates that listening (even for a native listener) is a very approximate art.

It could be claimed that using distractors in multiple-choice questions reproduces this process of forming and testing hypotheses. However, the problem is that these hypotheses originate, not in the recording itself but in the written items that accompany it. The result is that parsing in a conventional test imposes even heavier demands than parsing in a natural listening situation. The candidate not only has to form assumptions about the likely syntactic pattern that is evolving; but also has to form assumptions about whether the part of the recording being parsed is going to supply an answer to the current item.

5. Meaning construction

The third phase in making sense of an utterance comes closest to what is conventionally thought of as 'comprehension'; but again it does not consist of a single process. Test specifications often recognise differences between types of item:

- Local questions versus global questions;
- Extracting gist versus extracting detail;
- Extracting fact versus interpreting speaker intentions.

A different and more systematic way of thinking about these various levels of attention is provided by a psychological model (Field, 2008a) which distinguishes between three stages in developing the meaning of what a speaker says. They vary in the depth of comprehension demanded (and thus the cognitive demands on the test taker) and a tester can position questions at any or all of these three levels:

- A **proposition:** information from the recording at a very literal level. Here, the tester would ask for local factual information
- A **meaning representation:** where the listener relates a proposition to the context in which it occurs or draws conclusions which are not explicitly expressed. Here, the tester would ask about wider context or get test takers to draw inferences.
- A **discourse representation:** where the listener has to show that he/she has integrated everything that has been heard so far into a wider picture of what the recording is about - including speaker intentions etc. Here, the tester would ask about the relationship between ideas or get the listener to draw conclusions about the whole recording (including speaker intentions etc.).

5.1 Building meaning

A number of different processes assist a listener in developing a proposition into a meaning representation. They mainly reflect the need to fit the bare meaning of what is said into a context. If you hear *It's going to rain,* you cannot work out whether to reply 'That's good!' or 'That's bad!' unless you know if the speaker is worrying about a drought or planning a picnic. In adding to bare meaning, a listener might draw upon world knowledge, topic knowledge, knowledge of the speaker and situation or recall of what has been said to fair. He /she might also have to make inferences, supplying information that the speaker has not provided. For example, if one hears the sequence:

> *Bill lay on the floor. There was a knife beside the body.*

it seems reasonable to infer that the body is Bill's and that he has been murdered with the knife, though none of this is explicitly stated. In addition, a listener has to resolve references using terms such as *it, this, she, him, did so* etc. – terms that are often used more loosely in speaking than they are in writing.

This suggests quite a wide range of possible item types in listening tests, requiring candidates to:

- relate a bare statement to its context
- draw inferences
- resolve problems of reference
- interpret the speaker's pragmatics.

Testers quite often tap into one or more of these processes (in particular, they are quite responsive to parts of a recording that require inference). But they tend to do so relatively randomly, when the text of a recording suggests an opportunity. The result is that some tests are more representative than others of the full range of meaning building processes in which listeners engage.

5.2 Handling information

A further set of processes is used to integrate each new piece of information into a discourse representation, or picture of the recording as a whole. This final stage, where listeners have to decide what to do with the information they have obtained, is almost entirely neglected in current approaches to testing. It requires a number of decisions (Field, 2008a: Chap 13), which can be represented as:

- **Select:** the listener has to decide if a piece of information is important or not. If it is not, it can be allowed to decay.
- **Compare:** the listener has to check the reliability of a new piece of information by comparing it with what has been heard so far.
- **Integrate:** the listener has to add the new piece of information to the discourse representation, noting how it is linked to what went before.
- **Build a structure:** the listener has to build a line of argument based on major points and subsidiary ones (Gernsbacher, 1990).

The first and last of these processes in particular do not feature in most listening tests. The reason is that, with conventional test methods, the test setter decides which are the most critical pieces of information in a test and which are of low relevance, so candidates do not have to make these decisions for themselves. Similarly, the effect of asking a string of comprehension questions, of whatever type, is that the points addressed by the questions are treated as of equal value, and (except possibly in items testing global understanding) candidates are never required to organise them by importance or to trace relationships between them. Testers could argue that this is the price that one has to pay for reliability and ease of marking; but these information handling processes are important in many professional and academic settings, and their absence in standard tests is a serious challenge to cognitive validity.

6. An opportunity for local testing

It is, of course, possible to devise test methods and items that tap in to information handling processes. The more obvious ones include summary writing, oral report or an open-ended question that asks the candidate to list the speaker's (three/four) main points. It is also possible to use a 'table of contents' skeleton with gaps for main points and subsidiary ones to be filled in.

This type of exercise breaks with convention, and would not be acceptable in a high-stakes test with large numbers of test takers because of its unfamiliarity. But it is certainly something that could be contemplated by those setting local tests – for example, in-class tests of progress. All too often small-scale testing and the materials used in the teaching of listening emulate the formats employed in international tests, despite the fact that the conditions are very different.

The truth is that high-stakes tests have major constraints which prevent them from testing listening in a way that fully represents the skill. These constraints include the importance given to reliability and ease of marking, and (closely associated) an unwillingness to allow for individual variation or alternative answers. It is here that local tests administered to smaller groups possess a strong advantage. They are capable of involving a much wider range of listening processes by making use of innovative test methods. They can ask more open-ended questions, which enable them to test processes that are often unrepresented such as those involved in information handling. There is more scope for marking scripts on an individual basis; allowance can even be made for answers which are appropriate but differ from those anticipated by the test setter (Brown, 1995: Chap 1).

Local tests can also serve a 'testing for teaching' function. If testers at local level decide to design progress tests that focus on specific processes (decoding at word level, identifying word boundaries, parsing, relating meaning to context, inference, building a line of argument), the answers will enable them to identify areas of weakness in the listening behaviour of a particular group. These areas could then be practised by means of small-scale remedial exercises (Field 2008a). Even a simple policy of asking immediately after a test *'What answers did you give? Why?'* turns testing into a useful tool for diagnosing listening problems. In short, listening tests can indeed be formative and not just judgmental.

References

Brown, G. (1995). ***Listeners, Speakers and Communication.*** Cambridge: Cambridge University Press.

Field, J. (2008a). ***Listening in the Language Classroom.*** Cambridge: Cambridge University Press.

Field, J. (2008b). Revising segmentation hypotheses in first and second language listening. ***System, 36, pp. 35-51.***

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), ***Testing and Cognition***. Englewood Cliffs, NJ: Prentice Hall, ***pp. 17-30.***

Gernsbacher, M-A. (1990) ***Language Comprehension as Structure Building.*** Mahwah, NJ: Erlbaum.

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

16

Assessment for learning in listening and viewing - (based on Singapore's 2010 English Language Syllabus)

Tan Su Hwi

Abstract

Singapore's English Language Syllabus 2010 presents new directions for testing. Its emphasis on developing learners in multi-literacy and higher order thinking skills calls for a review in the function and form of assessment, which hitherto, has been the summative form of Assessment of Learning (AoL). This paper explores in particular, the integration of the listening skill with the viewing skill which the new Syllabus puts forth, and proposes an Assessment for Learning (AfL) framework. It also looks into the selection of print and non-print resources for testing, and the design of quality assessment questions for explicit teaching/testing of the integrated skills.

1. Introduction: An Overview of Singapore's English Language Syllabus 2010

Singapore's English Language Syllabus 2010 is an interesting progression from the former English Syllabus implemented in 2001. While keeping to the 2001 policy that English teaching should be learner-centred, the 2010 Syllabus takes into account the changing profile of the present school-goers; that these form a new generation of digital natives, many already comfortable with multimodal forms of communication, be it in the spoken, written, visual, gestural or spatial forms. All this translates into a few significant shifts in the English language classroom:

- Firstly, in terms of teaching material, the English Language curriculum in schools from 2010 onwards are expected to be enriched through the use of a variety of print and non-print resources, instead of being just text-book bound, as was the former practice. Print resources refer to physical artefacts such as newspapers, photographs and print advertisements. Non-print resources refer to digital resources such as web-based texts (e.g., online articles, blogs, wikis), CD-ROMs and DVDs, analogue resources such as films, TV and radio broadcasts, as well as live texts such as face-to-face encounters (e.g., conversations, interviews) and live performances (e.g., skits, puppet plays). The rationale behind incorporating authentic print and non-print sources at all levels is to expose students to texts with information-rich content, so as to promote the appreciation and use of English.

- Secondly, the receptive skill of listening is now integrated with viewing (see Appendix 1) and taken to be, "especially necessary in building a strong foundation in English at the start of language learning" (English Language Syllabus 2010, p.19). The Learning Outcomes for listening and viewing skills as specified in the Syllabus state that learners are to: 1) develop positive listening and viewing attitudes and behavior; 2) apply appropriate skills and strategies to process meaning from texts; 3) critically evaluate texts; and 4) construct meaning of a variety of extensive spoken, audio and visual texts.
- Thirdly, and as a follow-through from the above two mentioned Syllabus changes, the function and form of assessment has to be reviewed. Listening and viewing skills in particular, cannot be tested using the traditional summative assessment of learning, which was confined to audio-only testing for listening comprehension at the literal level of understanding. Now that learners are expected to be able to, "listen, read and view critically and with accuracy, understanding and appreciation of a wide range of literary and informational/ functional texts from print and non-print sources" (English Language Syllabus 2010, p.10), the design of assessment needs to be extended to engage students in different learner behaviour and increasing levels of critical literacy of multimodal texts.

However, up to the present time, the Singapore Examination and Assessment Board (SEAB) has yet to inform schools of how the listening and viewing skills will be tested. Often, teachers without assessment literacy or understanding of the new Syllabus are left to design their own tests for school-based assessments. Researchers have found that school-based assessments in general are of low authentic intellectual quality, focusing heavily on assessing students' memorisation of factual and procedural knowledge (Koh & Luke, 2009). In a research project undertaken in 59 Singapore schools (30 primary schools and 29 secondary schools) to examine the quality of teacher assignments and associated student work in 2004-2005, it was found that assessment practices by and large do not orientate towards students' understanding, let alone enhance learners' understanding (Tan, 2011). Critical metalinguistic skills of inference, evaluation and language appreciation are seldom modelled for students or tested. The consequent student work demonstrated a high level of reproduction of factual and procedural knowledge. This certainly does not sit well with the Learning Outcomes of the Syllabus 2010 which seek to develop students' higher order thinking and English literacy skills for real-world communication.

2. Unpacking Listening and Viewing Skills: A Case for Assessment for Learning

I propose here an Assessment for Learning (AfL) framework for listening and viewing skills, based on the recommendations set out in the English Language Syllabus 2010 (see Appendix 2). I agree with Tan's (2007) argument that assessment and learning do not simply relate together in causal or overlapping ways. Both are co-constitutive and dialectical - how assessment is constructed frames how learning exists and vice versa.

With a primary focus on the ongoing improvement for all students, teachers can use day-to-day classroom activities to involve students directly and deeply in their own learning in a formative fashion. The basic aim of AfL is to create for the individual student an understanding of the following processes (Chappuis, 2005):

- Where am I going? The teacher's role is to state, in words the students can understand, a clear idea of the learning target(s).
- Where am I now? The teacher's role is to train the students to be responsible for mastering the stated level of learning target(s). Through just-in-time, descriptive and regular teacher feedback, students learn to self-assess their progress.
- How can I close the gap? The teacher's role is to design lessons that revisit the key learning points of the lesson. Students are taught focused revision and self-reflection so that they can document and share with others their learning.

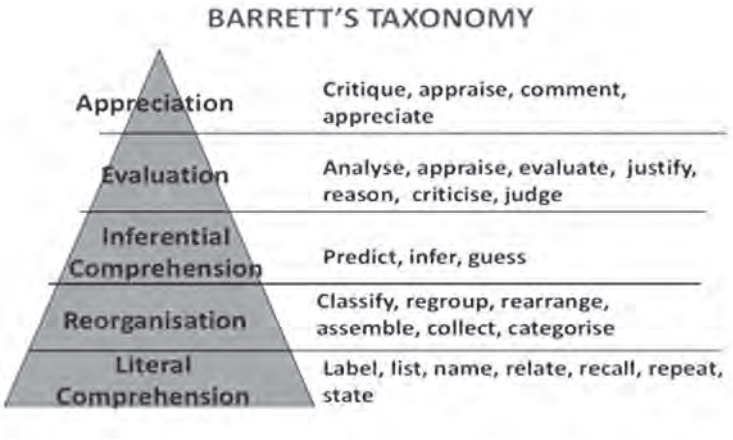
Inherent in the AfL model is a complex interplay of cognitive, affective and social aspects. Individual student's self-directed learning, peer-evaluation and group work all take centre-stage; and assessment is more about good teaching than testing in the traditional sense of the word (Davies, 2000). In AfL's design and practice, "the first priority is to serve the purpose of pupils' learning" (Black, Harrison, Lee, Marshal & Wiliam 2003, p. 2) where there is, "a classroom culture of transparency, strategic questioning by teachers and students, and an understanding of what is quality" (Yager, 2010).

3. Selection of Resources and Design of Assessment Questions

In the academic year 2010/2011, I taught the Language Testing and Evaluation module on the National University of Singapore's Master of Art (MA) in Applied Linguistics programme. Coincidentally, all the seven candidates who took my module were Singaporean teachers. Our discussion in class on AfL piqued me to design a prototype assessment for listening and viewing. Given "the importance of developing information, media and visual literacy skills" (English Language Syllabus 2010, p.16), I chose authentic materials from the internet: a) "an advertisement within an advertisement" video clip: the scene of a lady wanting to buy a burger in a wrong context is juxtaposed with the classic features of a prestige car; and b) a song with much angst about the world we live in and made famous by American idol, Adam Lambert, that students would readily identify.

To design strategic questions to engage learners' critical thinking, I used Barrett's Taxonomy of comprehension skills as a guide to measure the increasing higher order processing that has to take place. In summary, the categories of Barrett's Taxonomy are as follows:

- comprehending what is listened and viewed at the literal level;
- comprehending what is listened and viewed "in between the lines" (using reorganisation & inference skills); and
- comprehending what is listened and viewed "beyond the lines" (using evaluation and appreciation skills).


BARRETT'S TAXONOMY
Appreciation
Critique, appraise, comment, appreciate
Evaluation
Analyse, appraise, evaluate, justify, reason, criticise, judge
Inferential Comprehension
Predict, infer, guess
Reorganisation
Classify, regroup, rearrange, assemble, collect, categorise
Literal Comprehension
Label, list, name, relate, recall, repeat, state

Assessment for Learning Task 1

Aim:
To assess for students' level of listening and viewing comprehension

Instructions:
View the advertisement [http://www.youtube.com watch?v=eBPo0t69bi4] and answer the following questions.

Assessment for Learning Questions	Level of Listening & Viewing based on Barrett's Taxonomy
1. What did the lady ask for? ■ Books ■ Time ■ Food and drinks	Literal Comprehension
2. I can tell the reaction of the librarian from: ■ Her body language ■ Her tone of voice ■ Her gesture ■ The words she used	Reorganisation
3. How can you describe the librarian's reaction? ■ Disapproving ■ Shocked ■ Suspicious ■ Indifferent	Reorganisation
4. Why did the librarian react in this manner?	Inferential Comprehension
5. Did the lady understand the librarian's reply? ■ Yes ■ No	Inferential Comprehension

6. What was the intention of using a blond lady in the commercial?	Evaluation
7. Would the humour aspect be lost if the blond lady was replaced by: (i) a dark-haired lady (ii) a (blond) male	Evaluation/Appreciation
8. What connection does the commercial want to make between a blond lady and the Mercedes Benz car?	Evaluation/Appreciation

Assessment for Learning Task 2

Aims:
To enable students to listen and view information extensively; to enable students to discuss/debate with supportive statements what they have listened and viewed; to assess students' listening and viewing comprehension from the literal level to the appreciation level of Barrett's Taxonomy.

Instructions:

1. The class to listen to the song "Mad World" www.youtube.com/watch?v=bXGBWQdHsyQ (literal comprehension).
2. Students get in pairs.
3. Each pair provides a two to three sentence summary of what they heard (reorganisation skills).
4. Each pair share their summary with another pair and compare their answers (reorganisation skills).
5. Students listen to the song again, this time with lyrics flashed out in the video clip http://www.youtube.com/watch?v=RcmQvkojgz8

They note down the key words which provide information about the summary (literal comprehension).

6. Students compare the words in their groups.
7. Students view the music video with the theme of poverty attached to the song (inference skills). http://www.youtube.com/watch?v=yNaGr_biP6k
8. Students to reflect on how the music video presents the meaning of the song (evaluation and appreciation skills).
9. Each pair to discuss/debate with supporting statements of what they think the song means to them (evaluation skills).

4. Findings

I have used the two prototype tests on Singaporean teachers as well as teachers from the ASEAN region. Their feedback is that with appropriate scaffolding, appropriate questioning/probing techniques and modeling, students can be coached to acquire higher order thinking in listening and viewing multimodal texts. What they need is more curriculum time for AfL in practice.

5. Conclusion

In view of the increasing emphasis on self-directed and self-regulated learning as an indispensable 21st century competency, Singaporean policy makers, school leaders and teacher training agencies are considering incorporating AfL practice in their English language curriculum. The theme of the Ministry of Education's annual English Language Teaching Seminar in 2010 was "Assessment Literacies for the EL Curriculum" and its key-note speaker, Ms Karen Yager presented a paper focused on incorporating AfL into the teaching Syllabus. The English Language Institute of Singapore (ELIS) was also specially set up by the Ministry of Education in 2010 to provide English Language teachers with professional development. One of the flagship courses is assessment literacy. Policy rhetoric aside, the implementation of AfL would require stakeholders (parents especially) and school leaders to understand its socio-educational effects. Coming from an educational culture where summative assessment of learning is still a major driving force to grade, rank and certify learners, the taught syllabus can remain chained to large doses of passive learning and the drill-and-practice tradition of teaching. In fact, many teachers mistakenly equate AfL with "more mini-tests and mini-exams" – in order to "make up" for the high-stake year-end examinations which their schools have since lowered in percentage-weighting. This misnomer indicates to me that teachers themselves need to transit from the traditional "teaching students to test them" paradigm to the new "testing students to teach them" mindset; that AfL is not about grading students' ability (or the lack of it) to learn, but rather, relates to quality interactive teaching.

References

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). ***Assessment for Learning: Putting it into practice.*** Berkshire: Open University Press.

Chappuis, J. (2005). "Helping Students Understand Assessment". ***Educational Leadership, vol. 63, no. 3, pp. 39–43.*** Retrieved June 24, 2011 from http://teachingss.pbworks.com/f/Helping+Students+Understand+Assessment.pdf

Davis, A. (2000). ***Making classroom assessment work.*** Mcrville, British Columbia, Canada: Connections Publishing.

Curriculum Planning and Development Division, Ministry of Education (2010). ***English Language Syllabus 2010 Primary and Secondary***. Singapore.

Koh, K., & Luke, A. (2009). Authentic and conventional assessment in Singapore schools: an empirical study of teacher assignments and student work. ***Assessment in Education: Principles, Policy & Practice, 16(3), pp. 291 - 318.***

Tan, K. H. K. (2007). The Case for Qualitative Approaches to Assessment. In ***Alternative Assessment in Schools: A Qualitative Approach***. K. H. K. Tan. Singapore, Pearson Education South Asia.

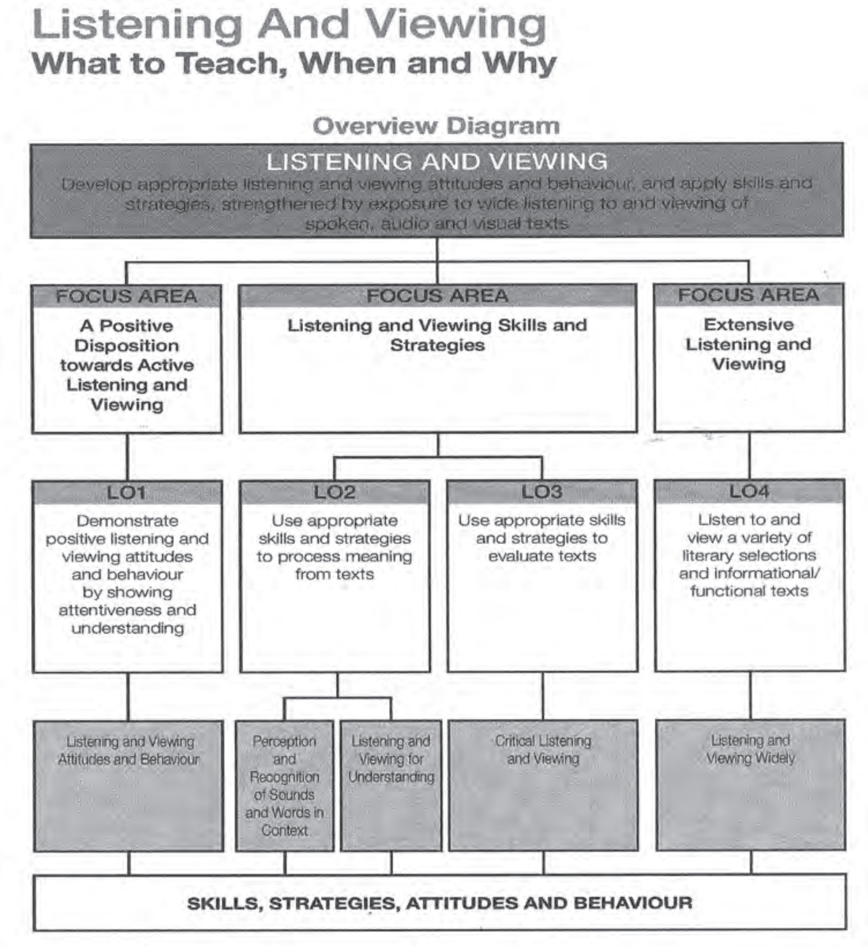
Tan, K.H.K. (2011). Assessment for learning in Singapore - Unpacking its meanings and identifying some areas for improvement. ***Educational Research for Policy and Practice***, 10,(2), pp. 91 -103.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), ***The future of assessment: Shaping teaching and learning, pp. 53-82.*** Mahwah, NJ: Lawrence Erlbaum Associates.

Yager, K. (2010). ***Building a Culture of Assessment for Learning In the English Language Classroom.*** Keynote address at the English Language Teaching Seminar, Singapore.

APPENDIX 1

Source: The Singapore English Language Syllabus 2010 (p.19)



Listening And Viewing

What to Teach, When and Why

Overview Diagram

LISTENING AND VIEWING

Develop appropriate listening and viewing attitudes and behaviour, and apply skills and strategies, strengthened by exposure to wide listening to and viewing of spoken, audio and visual texts

FOCUS AREA: A Positive Disposition towards Active Listening and Viewing

FOCUS AREA: Listening and Viewing Skills and Strategies

FOCUS AREA: Extensive Listening and Viewing

LO1: Demonstrate positive listening and viewing attitudes and behaviour by showing attentiveness and understanding

LO2: Use appropriate skills and strategies to process meaning from texts

LO3: Use appropriate skills and strategies to evaluate texts

LO4: Listen to and view a variety of literary selections and informational/ functional texts

Listening and Viewing Attitudes and Behaviour

Perception and Recognition of Sounds and Words in Context

Listening and Viewing for Understanding

Critical Listening and Viewing

Listening and Viewing Widely

SKILLS, STRATEGIES, ATTITUDES AND BEHAVIOUR

APPENDIX 2

The English Language Syllabus 2010 (p.19) is comprehensive in that it identifies for teachers how the listening and viewing skills are to be taught explicitly to learners:

- model positive learning attitudes and behaviour
- guide pupils in constructing meaning from a variety of spoken, audio and visual texts, beginning with the perception and recognition of sounds and words in context
- help pupils develop active listening and viewing skills - listen for details and listen for gist, make inferences, make predictions and listen selectively from Primary 1
- scaffold and model the learning of critical listening and viewing skills through the use of strategies and activities (e.g., brainstorming, concept-mapping, using pictures, tables, diagrams, conferencing)
- provide opportunities for pupils to listen to and view a variety of spoken, audio and visual texts for appreciation, enjoyment and personal development.

In essence, the above description is framed for Assessment for Learning, as teachers can build in the following key AfL strategies (adapted from Wiliam & Thompson, 2007) in the explicit teaching of listening and speaking skills:

- share learning intentions and expectations
- create effective classroom discussions, tasks and activities that elicit evidence of learning
- provide feedback that moves learners forward
- activate students in collaborative learning, reciprocal teaching, peer-assessment
- activate students as owners of their own learning.

17

Profiling graduating students' workplace oral communicative competence

Abdul Halim Abdul Raof, Masputeriah Hamzah, Azian Abd Aziz, Noor Abidah Mohd. Omar and Anie Atan

Abstract

Presently, there is no specific measurement for employers to gauge the true communicative ability of graduates entering the job market. There is thus a need to come up with a valid yardstick that would reflect the communicative ability of these graduates. This paper reports on research investigating how professionals at the workplace view and assess oral communicative ability of graduates. It discusses the process of validating the construct of the oral communication competence, and the process of developing and refining the competency profiles of graduating students. It also addresses other issues that may impinge on the assessment of communication skills.

Introduction

In an increasingly competitive and challenging global environment, workplace expectations are becoming more demanding. Current would-be employers are looking for more than mere academic qualifications. One attribute which has been recognised to be an important facet at the workplace is the ability to communicate orally (cf. Crosling and Ward, 2002). In countries where English is not the native language, more often than not, there tends to be an additional dimension to the oral communication ability. This involves the ability to communicate verbally in English. In Malaysia for instance, where English is deemed to be either a second language (Asmah, 1992) or a foreign language (Nunan, 2003), the ability to communicate verbally in English remains high on the list of employers. This can be seen from the results of a survey conducted by the Malaysian Employers Federation, which identified oral communication skills to be among the top skills sought by employers in new graduates (Star, 2011). Likewise, Briguglio (2003), conducting a study in a multinational company in Malaysia, also concluded that having good spoken English language ability is essential in order to perform work effectively. Workplace English language oral communication ability is thus a fundamental attribute that employers seek in would-be employees.

Despite this acknowledgement, there is hardly any valid yardstick used by employers which gauges the communicative ability of new graduates entering the job market. This then forms the impetus for this study. This paper discusses the process of validating the construct of the oral communication competence of graduating students devised through the collaboration of second language testing/teaching professionals and professionals in the workplace.

Development of Rating Scales in Assessing Workplace Oral Communication

A rating scale is "…a series of descriptors of certain criteria arranged hierarchically to show the differing levels of performance," (Abdul Raof, 2011). In formulating an oral communication rating scale, fundamentals involving what to assess, what criteria to use and how many criteria to include, could be addressed by referring to established literature on language performance (cf. Canale and Swain, 1980; Canale, 1983; Bachman, 1990). However, the formulation of a workplace oral communication rating scale is somewhat more complicated due to additional concerns which need further contemplation.

One concern pertains to real-world assessment criteria, which essentially is not related to the assessment of oral language proficiency per se. Jones (1985) in discussing the role and implications of nonlinguistic factors on performance-based language testing argues that on the one hand, some examinees who demonstrate substandard language proficiency may attain good overall scores due to their astuteness in certain areas such as personality traits. On the other hand, some examinees with high language proficiency may receive a lower score due to deficiencies in certain areas. Jacoby and McNamara (1999) believe that separating the linguistic criteria from the test context and content may lead to problems. This notion was made based on a study they conducted in which the language skills of Australian immigrant and refugee health professionals were assessed separately from their medical competence. Despite passing the test, Jacoby and McNamara reported that many of the test takers experienced problems during their actual clinical test due to poor English skills and lack of discourse competence. They thus argue that oral communication performance should not be assessed separately from professional performance as these two aspects are interrelated.

This then brings us directly to the next concern. Since oral communication performance is entrenched with professional competence, would it not make more sense for there to be some form of collaboration between language experts and workplace professionals particularly in the area of workplace oral assessment? In response to this concern, several studies on oral assessment have indeed involved both applied linguists and workplace professionals, with the latter group engaged in different capacities. In some studies, the workplace professionals served as co-raters with the applied linguists. The aim of such studies was to investigate the correlations between the two groups' judgments on the examinees' oral proficiency level (cf. Brown, 1995; Lumley, 1998). In other studies, applied linguists engaged workplace professionals as informants. Information obtained was then either used to aid in test construction (cf. Douglas and Selinker, 1993) or assessment criteria (cf. Douglas and Myers, 2000; Abdul Raof, 2004). This paper further extends the relationship between language specialists and workplace professionals by actively including the latter in the process of developing and refining graduating students' workplace oral competency profiles. It discusses the process of how workplace professionals' and applied linguists' view and assessment of graduates' oral communicative ability are incorporated in the formulation of graduating students' workplace oral communicative competence rating scale.

Process and Procedure in Developing Graduating Students' Workplace Oral Competency Profiles

This study involved a preliminary stage followed by a four-stage stepladder procedure adapted from Abdul Raof (2002). The procedure (see Figure 1) was adopted as its use has been shown to be useful in promoting active and continuous collaboration between two autonomous parties (cf. Abdul Raof, 2004, 2011), which in the case of this study involved applied linguists and workplace professionals.

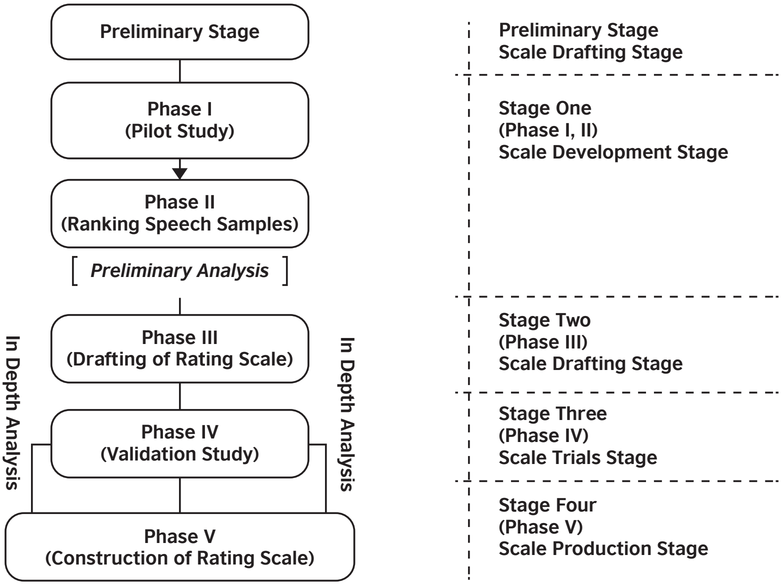
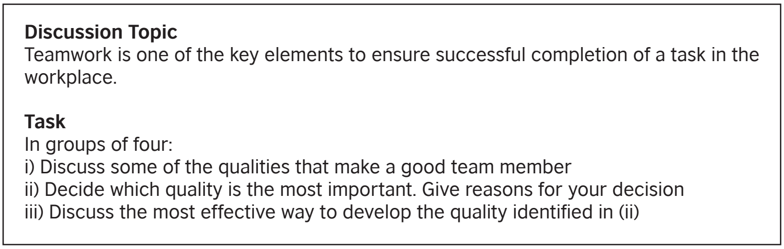


Figure 1: Procedure in Developing Graduating Students' Workplace Oral Competency Profiles (Adapted from: Abdul Raof, 2002)

The preliminary stage involved several meetings held among applied linguists involved in the study. The main aim of these meetings was to deliberate on the nature and criteria of graduating students' oral communication construct. This was then followed by the formulation of task sheets to elicit graduating students' oral workplace discourse. A sample of a typical task sheet is shown in Figure 2.



Discussion Topic
Teamwork is one of the key elements to ensure successful completion of a task in the workplace.

Task
In groups of four:
i) Discuss some of the qualities that make a good team member
ii) Decide which quality is the most important. Give reasons for your decision
iii) Discuss the most effective way to develop the quality identified in (ii)

Figure 2: Sample of Task Sheet Used to Generate Graduating Students' Oral Discourse

Topics chosen were to generally relate to the workplace. It was not possible to replicate authentic professional communicative event topics as the assessment scale is targeted at final year students who have yet to start work. It should be kept in mind that it is the students' current proficiency upon entry at the workplace that is to be measured.

Stage One involving the scale development stage then ensued. Using the discussion topics, several tape recordings of graduating students' oral communicative ability were produced. Each tape recording involved the participation of four graduating student volunteers, with different English language communicative ability. The tape recordings were initially viewed by applied linguists and the graduating students' oral communicative competence was then assessed and ranked without the use of any rating scale. The result was then compared to see any similarities or differences in terms of ranking and assessment. The tape recordings were then shown to a variety of workplace professionals involving human resource managers, general managers, directors, engineers, architects and IT Experts. They were also requested to assess and rank the graduating students' oral communicative competence, again without the use of any rating scale as it was the assessment criteria of the professionals that we were looking for. Based on their assessment it is worth noting that there was a disparity in the decisions made by the applied linguist and workplace professionals. The graduating students which the applied linguists rated to be the best in a particular group, were not rated as such by the workplace professionals.

Following the oral assessment exercise previously described, each respective participating workplace professional was then subjected to an interview session. This then formed the crux of Stage Two. Among the questions asked in the interview were:

- Who will you accept as an employee? Why?
- Why have you ranked Candidate X higher than Candidate Y?
  Why have you ranked Candidate Z the lowest?
- What other qualities would you like Candidate X (the successful candidate) to possess?
- What minimum qualities should a candidate possess to be considered for employment in your firm?

In Stage Two, responses generated by the workplace professionals were then analysed and the information was used to draft a rating scale. Once a consensus has been reached, a draft rating scale was devised. The draft rating scale comprised five criteria with six English language competency levels. The five identified criteria were *professional image, interactive ability, thinking ability, content,* and *language.* For the competency levels, the range was from 1, denoting *extremely limited* to 6, denoting *highly effective*, with level 3 considered to be a *functional level*. However, upon more analysis of the interviews and assessment made by the professionals, the applied linguists met and deliberated on the details of the scale. It resulted in the contraction of number of criteria from five to four, but with the competency level remaining unchanged at six levels. The four oral assessment criteria resulting from workplace professionals' feedback are *professional image, interactive ability, contribution to task,* and *language.* Figure 2 shows the development of graduating students' workplace oral communicative assessment scale.

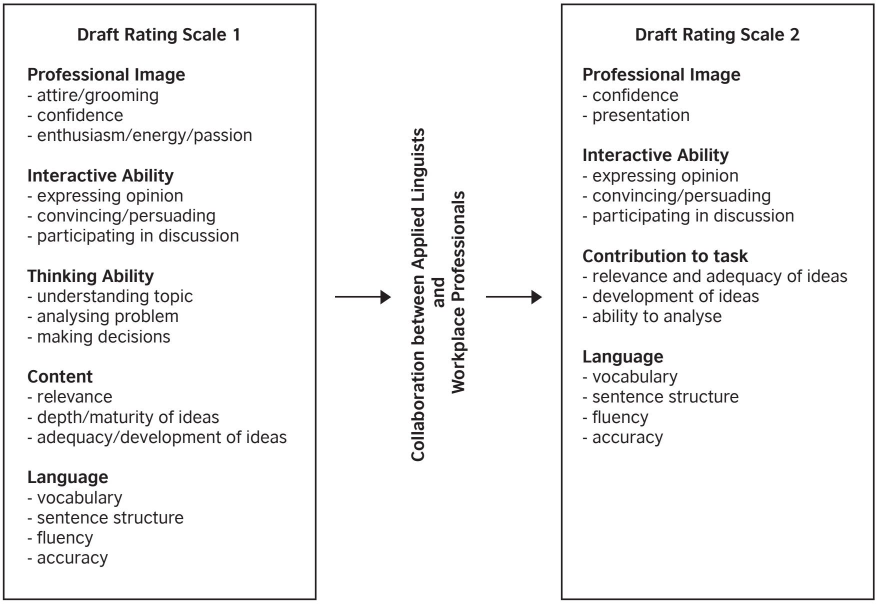


Figure 2: Development of Graduating Students' Workplace Oral Communicative Assessment Scale

In Stage Three, the refined rating scale generated in Stage Two was submitted to a validation process. Similarly, the validation process also involved the collaboration between applied linguists and workplace professionals. This was then followed by Stage Four, which is the final stage, involving the construction of the final, refined rating scale.

In conclusion, it could be discerned that the language element remains an assessment criterion in assessing graduating students' workplace oral communicative competence. Nevertheless, unlike applied linguists, who tend to emphasise the language aspect, workplace professionals tend to look at the 'whole package' with language being the vehicle for effective performance of tasks at the workplace. The oral assessment rating scale discussed in this paper is thus an example of how viewpoints of both applied linguists and workplace professionals can be operationalised.

Acknowledgements

This research is supported by the Research University Grant, Universiti Teknologi Malaysia (Grant No. Q.J130000.2611.04J91). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the position, policy, or endorsement of the funding agencies. The authors acknowledge the contributions made by other members of the research team: Anie Attan, Noor Abidah Mohd Omar, Khairi Izwan Abdullah, and Norzarina Sulong.

References

Abdul Raof, A.H. (2002). ***The Production of a Performance Rating Scale: An Alternative Methodology.*** Unpublished PhD thesis, The University of Reading.

Abdul Raof, A.H. (2004). In-roads into the Real World. In ***Language, Linguistics and the Real World (Volume II) - Language Practices in the Workplace.***

Abdul Raof, A.H. (2011). An Alternative Approach to Rating Scale Development. In B. O'Sullivan (Ed.), ***Language Testing: Theories and Practices.*** Palgrave Macmillan.

Asmah Hj Omar (1992). ***The Linguistic Scenery in Malaysia.*** Kuala Lumpur. Dewan Bahasa dan Pustaka.

Brown, A. (1995). The Effect of Rater Variables in the Development of an Occupational-Specific Language Performance Test. ***Language Testing. 12(1), pp. 1-15.***

Briguglio, C. (Ed.). (2003). Gathering linguistic data from two multinational companies: Intercultural communication in the workplace. ***In Proceedings from the 5th ABC European Convention.*** Lugano, Switzerland.

Canale, M. (1983). From Communicative Competence to Communicative Language Pedadogy. In J.C. Richards and R.W. Schmidt (Eds). ***Language and Communication.*** Singapore. Longman.

Canale, M. and Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. ***Applied Linguistics. 1(1), pp. 1-47.***

Crosling, G. and Ward, I. (2002). Oral Communication: The Workplace Needs and Uses of Business Graduate Employees. ***English for Specific Purposes 21, pp. 41-57.***

Douglas, D. and Selinker, L. (1993). Performance on General Versus Field Specific Test of Speaking Proficiency. In D. Douglas and C. Chapelle (Eds). ***A New Decade of Language Testing Research.*** Alexandria, VA. TESOL Publications.

Douglas, D. and Myers, R. (2000). Assessing the Communication Skills of Veterinary Students: Whose Criteria? In Kunnan, A.J. (Ed). Fairness and Validation in Language Assessment: Selected Papers from the 19th Language testing Research Colloquium. ***Studies in Languages Testing 9.*** Cambridge. Cambridge University Press.

Jacoby, S. and McNamara, T. (1999). Locating Competence. ***ESP Journal 18(3), pp. 213-241.***

Jones, R. L. (1985). Second Language Performance Testing: An Overview. In P. C. Hauptman, R. LeBlanc, and M. B. Wesche (Eds.), ***Second Language Performance Testing.*** Ottawa. University of Ottawa Press.

Lumley, T. (1998). Perceptions of Language-Trained raters and Occupational Experts in a Test of Occupational English Language Proficiency. ***English for Specific Purposes 17(4), pp. 347-367.***

Nunan, D. (2003). The Impact of English as a Global Language on Education Policies and Practices in the Asia-Pacific Region. ***TESOL Quarterly 37(4), pp. 589-613.***

Star (2011). English Works. ***Star***. April 10.

PROGRAMME EVALUATION

18

Issues in the evaluation and assessment of education projects and programmes: The relationship between donors, governments and stakeholders

Philip Powell-Davies

Abstract

It is a matter of good practice to monitor and evaluate programme interventions and to ensure this is incorporated into the design of projects. This is especially important in large-scale donor-funded projects and programmes which work with and for host governments, and are subject to increasing pressure to be accountable to government for the tax-payers' money that is spent on development aid. This study is based on the author's experience in leading and observing evaluations of donor-funded education and English language projects in Asia, the Middle East, E. Europe and Africa and offers an opportunity to pull back and reflect on a number of issues that this experience has identified. This study is as much a story of the importance of relationships as it is an analysis of tools and standards of project and programme management.

Introduction

There has been a renewed interest in impact evaluation in recent years amongst development agencies and donors. A 2006 Center for Global Development (CGD) report called for more rigorous impact evaluations, where 'rigorous' was taken to mean studies which "tackle the selection bias aspect of the attribution problem" (Holland, 2007; CGD, 2006). This argument was not necessarily well-received in the development community; partly due to the mistaken belief that supporters of rigorous impact evaluations were pushing for an approach solely based on randomised control trials (whereby study subjects, after assessment of eligibility and recruitment, but before the intervention to be studied begins, are randomly allocated to 'receive' or participate in one or other of alternative interventions). In fact, the CGD report argued for a broad use of many methods to evaluate impact, both in qualitative and quantitative terms.

There are a number of reasons which explain both the renewed interest in evaluation among donors and its role and significance in complex projects. The most important of these include:

- Accountability (to government, tax payers, stakeholders and beneficiaries) and empowerment of recipient communities;
- Desire for better cooperation between development partners – donors, governments, managing agents and beneficiaries;
- Developing an evidence base from which policy makers can make decisions;
- Need to measure results and often use numbers to do that with a focus on **outcomes** (what is being achieved), **outputs** (what is being produced) and **inputs** (how the money is being used);
- Improving project performance by building a synthesis of qualitative and quantitative data, combining so-called standard indicators and a meaningful narrative about project processes, stakeholders' interests and the wider socio-cultural context;
- Adopting new ideas in project design, activity and implementation;
- Building a case for enhanced financial support;
- Documenting and publicising project achievements to date;
- Enhanced project management capabilities - moving from efficiency of implementation to effectiveness of project interventions;
- Encouraging and developing a culture of learning within a project;
- Demonstrating Value for Money (VfM) particularly in quantitative terms;
- Fulfilling international agreements on education and development, such as the Millennium Development Goals.

We can see from this list that there are potential tensions between donor and recipient perspectives of evaluation, as well as stresses between several often conflicting aims which are packaged into quite simple evaluation missions. Evaluation may be motivated by a desire to prove the worth of a given intervention, which implies that external funding for evaluation is strongly demand-driven. This then favours those projects expected to have benefits by their advocates. After all, why would someone commission an evaluation if the results were expected to be negative? The counterbalance to this more 'promotional' aspect of evaluation is a commitment to the independence of the evaluation study and its findings supported by input from strong governments/other stakeholders who are intelligent customers of more and better evaluation.

Different donors disbursing development assistance follow their own procedures for periodic review of project and programme activity, and most institutions (typically but not exclusively governments) who are the recipients of such development assistance are frequently the subject of donor-funded reviews and evaluations themselves. A large amount of money is spent on this aspect of project activity (typically in the region of 10% of total project spend) but comparatively little is known about the efficacy of all this effort; how it is used for the future shape of the project; the degree to which it represents a 'tick-box' mentality by donors who may have little commitment to following through on the evaluation findings; and/or the extent to which it contributes to the achievement of overall developmental goals of quality education.

Whose project is it anyway?

The data derived from different stakeholders in evaluation exercises clearly reflect their differing interests and this in turn raises a number of questions with respect to:

- Degrees of participation – who is involved; what is the nature of the participation – is it merely consultative, more formative or transformative?
- Degrees of ownership – to what extent does a recipient own a project; how do individuals and communities demonstrate ownership, and how is this validated?
- Relative perceptions of the value of evaluation among stakeholders;
- How the findings and results are used to reconfigure the project, amend project goals and activities with sensitivity to the context in which they are operating rather than simply reflecting the hyper-rationality of the project planning frameworks;
- Lessons learned from both the process and the outcomes.

Formative and summative approaches

Project evaluations typically include both formative (mid-course) and summative (final) instruments. Ex post evaluations, intended to occur sometime after a project has finished, and aimed at investigating longer-term impact are rarely seen despite the many recommendations that are made in evaluations for this to be done. This is an opportunity missed and fails to understand the long-term nature of change in education systems, practices, behaviours and so on.

Probably the most important time to conduct a project evaluation is approximately 2/3 years after its start (depending on the length of the project) as part of a formative approach. This is a widely used mechanism for reviewing progress against original objectives, identifying bottlenecks, making mid-course changes and re-allocating budgets. Contrary to their name, mid-term evaluations often take place in the latter half of the project cycle, closer to the finish of the project. This is because most projects have a start-up time of about one year, during which teams and systems are set up. Given then their timing in the project cycle, mid-term evaluations frequently double up as a practice run for second phase funding or as a justification for terminating contracts and partnerships.

Final evaluation, though written into the design of many large projects, is less frequently seen in practice. Final evaluations that do take place are those situated in the middle of two successive project phases. Since there is an implicit competition for resources it is perhaps understandable that these "are invested in projects which have a future rather than simply those with a past" (cf. Bajaj 1997). Almost all projects, however, do require some form of internally-generated project completion report to be compiled but the detail, quality, timeliness and degree of learning shared from these reports varies greatly across agencies.

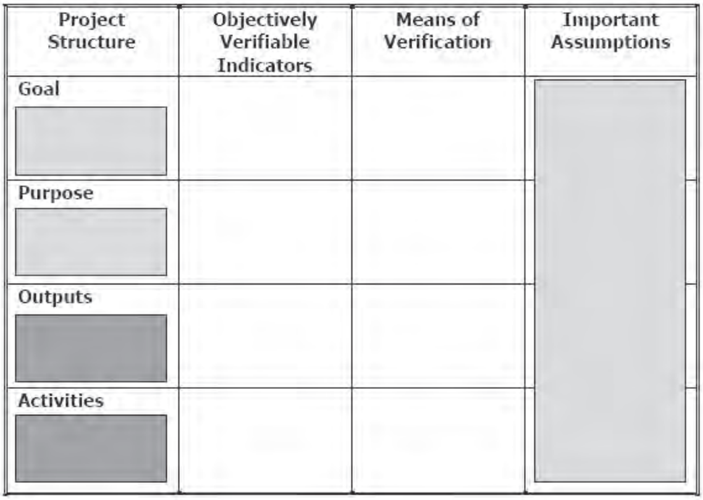
The logical basis of project planning

Formative assessment approaches, often called Output to Purpose Reviews (OPRs), clearly have strong potential in theory to influence the future conduct and direction of a project intervention. But let's first examine where the 'output' and the 'purpose' fit in the Logical Framework hierarchy (the logical framework is a common analytical tool used to plan, monitor, and evaluate projects. It derives its name from the logical linkages which connect a project's means with its ends).

The core of the Logical Framework is a temporal logic model that runs through the matrix. This takes the form of a series of connected propositions:

> *If XXX Activities are implemented, and XXX Assumptions hold, then XXX Outputs will be delivered*
>
> *If XXX Outputs are delivered, and XXX Assumptions hold, then X Purpose will be achieved.*
>
> *If X Purpose is achieved, and XXX Assumptions hold, then X Goal will be achieved.*

These are viewed as a hierarchy of hypotheses, with the project/programme management agency being responsible for the validity of hypotheses beyond the output level (see Figure 1).



Project Structure	Objectively Verifiable Indicators	Means of Verification	Important Assumptions
Goal			
Purpose			
Outputs			
Activities			

Figure 1 – The main components of a logical framework matrix.

Stakeholder involvement in reviews

OPRs are typically carried out by small teams of independent consultants (3-5 people, depending on the complexity of the project). Such reviews last between 1-4 weeks in duration. Stakeholder involvement in the process varies considerably, but as a principle it is important because it:

- Builds in contextual knowledge and relevance;
- Empowers relevant groups of people involved in the project intervention; and
- Strengthens individuals and institutions in-country.

Levels of involvement are influenced by a number of variables: who is funding the review; what is the focus of the review; and what will happen to the findings.

Planning Stage

There is no standard procedure for scoping the review. Issues are often identified in an ad hoc way, though this is increasingly in partnership with stakeholders and/or derived from internal monitoring processes, quarterly and annual reports. This stage is sometimes not used at all and subsumed into formulating a set of Terms of Reference (ToRs) for the review itself.

Terms of Reference
These are almost always written by the donor agency. It is often very difficult to get input from a host government on a draft set of ToRs. There are numerous reasons for this and the one most often cited by donors is the low level of capacity in host governments. This may not always be the case, however. Projects which are closely guarded by donors and perceived to be driven by them often exclude the input of recipient governments, for dubious arguments about objectivity. And this is also the case when projects are contracted out to managing agencies who are often closer to the donor and take direction from the donor rather than the host government for whom the project is intended. A degree of incoherence in this scenario should be obvious to the reader.

Choice of consultants
The choice of consultants is made by the donor and as the funding is provided by the donor this strengthens their rationale for selecting the consultants. The timing of the review is usually a consultative process sorted out between the donor, the managing agency and other stakeholders.

Conduct of the review
When we think of the way in which the review itself is carried out, briefing of the consultants is done by the donor (sometimes) in conjunction with the managing agency and very occasionally the recipient host government. Logistics is handled by the managing agency. The review methodology is the responsibility of the consultants and routinely signed off by the donor. Similarly, background data sourcing is done by the consultants, though managing agencies are required to provide internal project documentation for review.

As far as reporting is concerned, a detailed discussion of findings (debriefing) session is usually conducted by the consultants for the donor. This is widely regarded as important and participation by senior stakeholders is usually high. Occasionally donors restrict the debriefing, and managing agencies are not invited. Draft and final reporting is written by the lead consultant and provided to the donor, who in turn is then responsible for sharing the findings with host governments and agencies. This can be a long and tedious process if an adequate debriefing is not carried out. Major challenges are posed to consultants, as donors are increasingly uninterested in detailed review reports. In recent months the author was requested to provide a 6-page OPR report to capture the complexity of a large, multiple output project valued at over 50m GBP.

Conditions for success in project evaluation

It is possible to summarise some of the key success criteria in project evaluation:

- Recipient willingness and cooperation – this cannot always be presumed especially in government circles. This can be due to capacity issues, lack of adequate briefing by managing agencies and donors, ‘distance’ from the project, and/or pressure of work;
- Adequate resources – both financial and human. Expertise to conduct evaluations and coordinate complex processes is always at a premium;
- There are varying degrees of comfort with participatory approaches to evaluation and impact assessment. Donors are often suspicious of it. And in turn, recipients are often suspicious of donors’ motives and interests. Weak governments can be steam-rollered into accepting approaches and solutions which they frequently do not fully understand;

- Donors need to work with on-going monitoring and evaluation processes. Effective projects incorporate such processes as specific outputs in their logical frameworks;
- Knowledge sharing and good partnership between donors, managing agencies and recipients (and within large complex projects comprising multiple partners who may have different aims and working practices);
- Focusing on cross-cutting enablers which can help to build the case for funding English language projects – rights, gender, equity, social inclusion, political economy of education etc. Frequently these are the issues that tend to be sidelined by otherwise technically strong projects which focus their energies on technical inputs to e.g. teacher training curriculum development and so on. A narrow 'technicist' approach is unlikely to enable a project to achieve its longer-term sustainability agendas; and
- Clarity about sustainability and institutionalisation in order that donor-funded initiatives are appropriately positioned.

What does this tell us about project evaluation?

Project evaluation should be viewed in a positive light despite the fact that stakeholders will have different agendas – eg. donors are always tend to be interested in accountability; project teams with learning and endorsement of their approach. Evaluation brings benefits in allowing project teams to go beyond their day to day routines and reflect on why they are doing what they are doing, what effect it is having and what they need to adjust to achieve their goals. Evaluation also brings fresh perspectives to bear and exposure to new ideas and concepts as well as promoting independent documentation and verification of a project's work.

However, donor-funded evaluations tend to be confined to a narrow project cycle mode and are often motivated by management control of outputs with less focus on process, partly due to the fact that processes unfold slowly over time and their influence requires patience and longitudinal study to understand them in detail.

The extent to which evaluations can be led by a project agency or beneficiary is in doubt when the donor funds the exercise. VfM issues reinforce the agenda and tend to impose conditions which pay little heed to the specific context in which a project is operating and ignore the needs of the recipient stakeholders of the project.

Evaluations tend to work within the overall scope and design of the existing project and rarely result in a revaluation of basic approaches and premises. This imposes severe constraints on the degree to which the evaluation can address fundamental flaws in project design and purpose. In particular few donors are interested in allowing major changes to logical frameworks once they are agreed. The UK's Department for International Development (DfID), for example, will not consider amendments to be made to a project goal or purpose in an OPR review, even if it is inappropriate and identified as unhelpful by an independent evaluation team. This may make life easier in the short term but in the long term such a position hampers the ability of a project to achieve its goal. A shared approach from donor, government and agency can produce a more holistic picture through enhanced participation and help to support integrated action planning due to better understanding of both the process and findings.

Where do English Language (EL) projects fit into the picture?

EL projects have in recent years been neglected by donors for funding based largely on a thinly explored argument that EL interventions privilege elite communities and do not contribute to alleviating poverty. Using a more inclusive social development lens in recent years there are signs that EL projects are once again on the table as part of an integrated approach to education development aid. DfID currently funds a 50m GBP project in Bangladesh whose purpose is 'to increase significantly the number of people able to communicate in English to levels that enable them to participate fully in economic and social activities and opportunities'. While it is very encouraging to see EL projects being deemed worthy of support as part of a development agenda, the wording of the purpose of this project is problematic especially when we consider that the outputs of the project do not specifically have economic and social activities and opportunities written into them. This in turn means that a project designed in this way has no mandate and little ability to work with relevant ministries who control labour market policy or economic planning and development. And, in turn, the possibilities of even an efficient and well-run project being able to clearly demonstrate that it is working towards that purpose is hampered by a narrow technical focus on EL methodology, training courses and curriculum innovation and so on. This is not uncommon and has been observed in other EL project, although smaller in scale. There is a clearly a mismatch between the formulation of a project's purpose and the design of its logical framework such that it is designed appropriately to achieve the purpose that has been identified for it.

It is not impossible to build a case for the relevance of EL projects in a development context and for donors and governments to help fund them. The key principle of review methodology is to gather data to base observations and recommendations on evidence to make a coherent case related to the linkages to wider government reform agendas, building levels of awareness and interest among wider stakeholder groups; showing the relevance of child-centred communicative approaches to better learning outcomes; clarifying thinking about what constitutes the 'reach' and 'impact' of a project (see Figure 2), as well as sustainability and institutionalisation which frequently hamper the ability of projects to demonstrate their value and worth. All of these elements will then enable projects to understand issues of social inclusion, labour market strategy, political engagement and so on and align their interventions more appropriately.

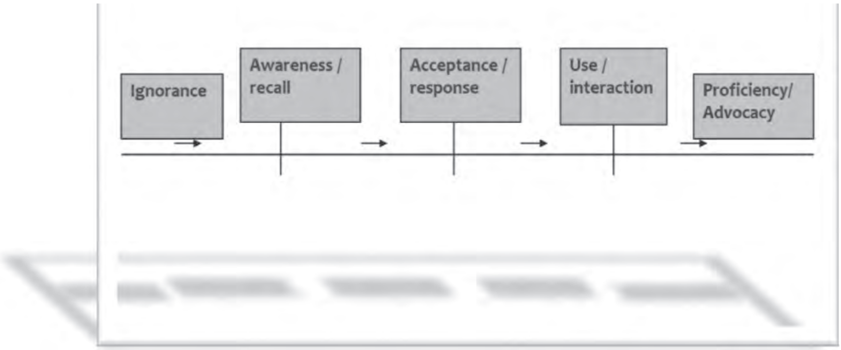


Figure 2 – Project reach and impact: the customer journey along a route of more frequent and intense interaction with project interventions, from ignorance to proficiency and advocacy. (Source: Mohun, A. 2011).

EL projects are sometimes accused of being remote from and outside the education mainstream but as we see from many of the articles in this publication, governments are increasingly considering introducing English at a younger age, are experimenting with teaching maths and science through English and investing large sums of money in rolling out national programmes of EL innovation. If the relationship with a donor can help to galvanise those developments then projects need to be designed to make outputs reflective of the wider context in which EL innovation take place – i.e. if EL curriculum reform is planned then how is that related to wider and on-going reform projects within government (e.g. sectoral or cross-sectoral national strategic plans) and how can it be linked in to them? Actually what we are talking about is exploiting opportunities through advocacy strategies for wider understanding of what is being aimed at; capturing best practice through case studies and policy papers to influence thinking at the strategic level; commissioning studies that enable a project to understand the socio-economic context in which it has to operate and also to set the agenda; developing partnerships with government, NGOs and academia and the private sector; ensuring that EL innovation in pedagogy fits the cultural context of the country and community; and crucially, understanding social inclusion issues better so that the role of English in creating opportunities for poorer communities is clearly understood and communicated.

This is likely to achieve a much more integrated approach with as many stakeholders as possible working together so that evaluation becomes ‘doing with’ rather than ‘doing to’.

Conclusions and Lessons Learned

Most national and donor systems of monitoring and evaluation are concerned with the progress of implementation, rather than assessing the social, economic and environmental impacts of projects. Also, there seem to be few systems that assess the impact of policy interventions emerging from macro-level measures to link education reform and EL innovations, such as privatisation, rights-based approaches, and so on. In developing countries, donor agencies have played a role in planning, implementing and financing various socio-economic development programmes and projects. In many cases, the outcomes of these interventions do not match the intended objectives. It has been argued that due to the lack of on-going evaluation many governments fail to learn, in time, the way a project is unfolding and the manner in which it is generating benefits. There are also many who simply do not see the benefits of evaluation and consider it to be a donor-driven activity of little governmental management use. Donors are increasingly investing in evaluation capacity building activities for government decision-makers. The success of these initiatives seems to have been constrained, among other things, by the lack of a unified approach among donors; inadequate appreciation of governmental culture; confusion about concepts and methodologies; lack of long-term commitment; and lack of either interest or resources, or both, from the recipient governments (Garbarino and Holland 2009). Future evaluation capacity building work will need to make a careful analysis of these constraints and approach the subject with far greater sensitivity and technical knowledge.

Both donors and recipients feel that evaluations make a positive contribution and result in value addition to their work. However, donors and recipients view the gains from evaluation very differently. For donors the evaluations are useful for ensuring the accountability of their investments and for improving project management. For recipients they are useful in that they create a space for reflection and stock taking and provide them with a fresh perspective on their work. Most donor evaluations are designed in consonance with the donors' needs and therefore take project objectives as the starting point. Learning from project evaluations could be made more pertinent from the recipient's angle if the evaluation expanded its brief to re-interpret and review the recommendations additionally from the perspective of the organisation's overall goals and strategies.

By and large the process of evaluation is managed by the donor and by external consultants with recipients having less say in the planning and conduct stages of evaluations. Enhanced consultation does take place in the reporting stage via de-briefing sessions and eliciting comments on draft reports but this occurs too late and fails to build a sense of ownership and commitment to the utilisation of the evaluation on the part of the recipient. Modifications to the process, shifting the responsibility for some components (e.g. drafting terms of reference), ensuring greater consultation and transparency in others and instituting mechanisms which allow differences of opinion to surface early and be dealt with before finalisation could enhance the utility and acceptability of evaluations.

Though the bulk of evaluation practice is in the conventional non-participatory mode, several donors in partnership with their recipient counterparts are innovating and moving in the direction of more participatory approaches. These approaches encourage a greater degree of utilisation of evaluation results and a very positive contribution to institutional strengthening in the recipient agencies.

A large number of internal review mechanisms are in place in most recipient agencies. Sometimes it makes better strategic sense for the donor to draw upon these instead of instituting separate evaluations.

Several constraints hamper the emergence of participatory evaluations as the norm in donor funded evaluations. Some of these are recipient unpreparedness, shortage of experts and facilitators and lack of support for the concept within donor bureaucracies. Moreover, participatory evaluations, more demanding of resources, may not be necessary in all cases.

For most donors the way forward will comprise a combination of approaches: evolving participatory approaches, making necessary changes in conventional evaluation modalities to better reflect recipient concerns, and selectively supporting those internal review processes in the recipient agencies which can simultaneously address donor needs.

References

Bajaj, M. (1997). ***Revisiting Evaluation: A Study of the Process, Role and Contribution of Donor funded Evaluations to Development Organisations in South Asia.*** Ottawa: IDRC.

Center for Global Development. (2006) ***'When will we ever learn? Improving lives through impact evaluation.'*** Report of the Evaluation Gap Working Group. Washington, D.C.

Cummings, R.J. (1996). ***"Making Evaluation Count"***. Paper presented to the 1996 Australasian Evaluation Society (Western Australian Branch) Conference, Perth, 18 September 1996.

Garbarino, S. and Holland, J. March (2009). ***Quantitative and Qualitative Methods in Impact Evaluation and Measuring Results Issues Paper.*** DfID Governance and Social Development Resource Centre. Social Development Direct

Mohun, A. (2011). Unpublished Output to Purpose Review Report, commissioned by DfID Bangladesh.

Preskill, H. (1994). “Evaluation’s Role in Enhancing Organisational Learning: A Model For Practice” ***Evaluation and Program Planning 17(3), pp.291-297.***

Prowse, M. (2007). ***Aid effectiveness: the role of qualitative research in impact evaluation.*** ODI Background Note. December 2007.

19

Programme evaluation: Interconnectivity of variables

Kyungsook Yeum

Abstract

Teachers have been exposed to various types of development opportunities. However, the investment that teachers put into professional opportunities is not necessarily always rewarded. In this paper, a theoretical framework will be outlined that can be used as a model to evaluate an INSET programme in similar settings. While adopting a 'process-oriented' evaluative approach, we can observe several dimensions of the programme to explore how the main stakeholders and context variables interact to determine the quality of the programme. The rationale for a synthesised approach to programme evaluation to look at those variables will be discussed in the following sections.

1. INSET: Curriculum and Evaluation

Even though it requires a long time to improve teaching, ways of strengthening the INSET system have frequently been proposed. We can borrow an idea from Stronkhorst and Akker's (2006) recommendation to improve science education in Swaziland. Short-term and long-term effects (before-during-one year after) of an in-service intervention resulted in changes. They concluded that a more flexible approach that combines realistic outcomes for specific target groups in specific situations with an appropriate design has a greater potential of success. Walters (2006) also emphasises the, "importance of course-based training" and "sufficient schools based support" to maximise the potential for teacher learning.

An INSET curriculum designed to address teachers' needs in a specific teaching context must be one of the keys to achieve quality outcomes. Guskey (2003) concludes that the effectiveness and impact of professional development is complicated and complex, and therefore it should be contextualised:

> *"Because of these powerful contextual influences, broad-brush policies and guidelines for best practice may never be completely accurate. Still, by carefully considering these contextual elements and making decisions based on specific evidence of student learning, visionary school leaders can better ensure that their professional development programs and activities will meet with success."* (Guskey 2003, 16)

Hayes (1996) also expresses the imperative to consider the school context and the teachers situated in it to develop a successful INSET for them.

As for designing a quality INSET programme, meeting the specific needs of teachers should result in a desirable, relevant outcome. It is not an easy task to plan 'backward,' as Guskey (2001) suggested, to think about students learning outcomes first, and reflect the needs in the design of the INSET curriculum to enhance the targeted training. However, addressing teacher trainees' own needs must be fundamental to enhance the quality of the INSET programme.

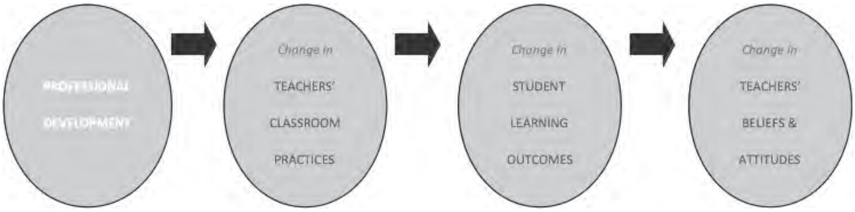

PROFESSIONAL DEVELOPMENT
Change in
TEACHERS' CLASSROOM PRACTICES
Change in
STUDENT LEARNING OUTCOMES
Change in
TEACHERS' BELIEFS & ATTITUDES

Figure 1 - A model of teacher change (after Guskey (2002), 49).

According to the model, any noticeable change in teachers' attitudes and beliefs will be observed only after they confirm any new instructional approach or practice works in their classrooms. As for needs analysis, traditionally, the questions are asked about what the trainees' need and what and how it should be dealt with (West 1994). Figure 2 shows the perspectives involved to determine the nature of curriculum and course design.

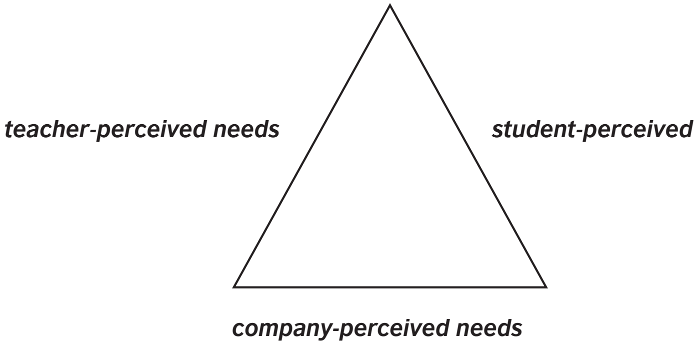


Figure 1 - Needs analysis perspectives (West 1994, 6).

In addition, a few research studies have been done using the tripartite perspectives that consider the needs of teachers, students, and administrators. Kikuchi's (2004 and 2005) research shows that there is a difference in perception about learners' needs among administrators, teachers, and learners. The implication is that using multiple information sources is needed to find the real needs to be addressed in a quality curriculum. Different perceptions lead to certain attitudes and motivations, and then eventually to teaching/learning behaviours in the educational context. "The importance of perception and its influence on attitudes and subsequently on behavior cannot be understated," (Morrison 1998, Editorial).

2. Process Evaluation of Programme Quality

In the current evaluation model, the process-oriented approach is adopted as a way to evaluate a programme's quality under a general belief that a quality process would eventually lead to sustainable quality outcomes (Payne 1994; White 1995; Fitzpatrick, Sanders and Worthen 2003; Royse, Thyer, Padgett and Logan 2006).

In the evolutional history of programme evaluation, there has been a focus shift, "as a move away from a concern with tightly controlled experiments focusing on the analysis of product, or student achievement, to a concern for describing and analysing the process of a programme," (Lynch 1996, 39). Accordingly, answering the "how" and "why" have become important parts of programme evaluators' and stakeholders' roles in the evaluation process.

Process-based evaluations are geared to fully understanding how a programme works to produce a certain result. Process evaluation can show how a programme is working, as well as whether it has achieved any quality standards. Long (1984) argues that,

> *"Process evaluations offer many benefits for teachers and administrators alike. Of these, the most important is that they can document what is actually going on in classrooms, as opposed to what is thought to be going on,"*
> (Long 1984, 422).

In addition, White (1998) deepens the meaning of the term by involving people (stakeholders) within the context and also by recognising the value of process quality - "it is process quality and effectiveness that lead to sustainable quality outcomes," (White 1998, 137). A higher priority should be placed on the process itself in which different variables affect each other and collaborate to determine a certain level of quality (Crandall 2000).

3. Interconnectivity of Quality Drivers in the Programme Context

3.1. Organisational Culture and Educational Quality

In this evaluation model, 'quality' is used as a term to define the process quality in which four major agents interact with each other: educational officers; programme administrators; teaching staff; and trainees. The idea of a desirable organisational culture (Davidson and Tesh 1997, Bennis and Nanus 1997, Senge 1990, and Liebowitz 2008) is directly applied in the concept of a good educational institution. A sharing culture can gain trust from all stakeholders. Accordingly, gaining support from the members to achieve shared goals is more probable. Among the indicators of the quality of an educational institution stated by Morris (1994), the majority of them are characteristics related to the organisational culture and administration.

Another notable dimension that should be discussed in the organisational culture, particularly in the language teaching organisation, is the potential conflict that could make communications among participants difficult. As White et el (2008) clearly points out, a language teaching organisation is potentially 'bi-cultural.' The academic and administrative sides of a language programme administration could show different sides of the overall organisational culture. If we consider the socio-cultural aspect of classrooms and the chemistry of students to influence quality of a programme, it will be 'tri-cultural'.

3.2. Interconnectivity of Quality Drivers

In connection to quality measurement, Tam (2001) summarised a few common approaches to quality assurance in higher education. Those models place students at the centre of evaluation and their learning outcome or their learning process itself becomes the focus of evaluation. When we evaluate students' entire learning experience, other context variables will also be automatically involved.

On the other hand, teacher effectiveness/teaching quality is definitely one of the major areas in determining the quality of a programme (Pennington and Young 1989; Freedman 1989a; Freeman, Orzulak and Morrissey 2009). While measuring school effectiveness, teacher effectiveness has been singled out as the central indicator of a programme among school variables (Ellett and Teddie 2003). In that case, teacher beliefs, principles, and practices were regarded as the main indicators of programme quality. Again, even in evaluating a programme from that particular point of view, students and their learning outcomes cannot be ignored.

Crabbe (2003) and Payne (1994) support a comprehensive evaluative approach. Morris (1994) also provides an evaluative framework that includes several characteristics as indicators of the quality of a school or educational institution. Accordingly, the evaluation of one specific area in a language programme still requires a more comprehensive approach to involve other factors (Tuffs 1995; Rea-Dickins 1994; White 1998; Ellett and Teddlie 2003; Fred, Newmann, King, and Youngs 2000; Freeman 2009; Bailey 2009).

Even though little research has been done to show a clear connection between those areas, the interconnectivity of different variables has been implied or suggested through research on classrooms in the educational field. In this regard, Kiely and Rea-Dickins (2005) acknowledge a wide scope of evaluation for a language programme for a formal evaluation study. They particularly point out the interrelated aspect of diverse human and programme factors, and also the connectivity of processes and outcome:

> *"The evaluations – assessments, audits, inspections – which generate these judgments suggest a strong role for users and programme participants: they are stakeholders whose experience of the programme is the key to unlock the 'black box' of quality....The challenge of evaluation processes engaging with notions of quality is to capture, in a credible manner, the drivers of quality and the factors which mediate them."* (Kiely and Rea-Dickins 2005, 11)

The critical term, "the key to unlock the 'black box' of quality," implies particularly a couple of things. First of all, the term, 'black box,' signifies the complex and complicated nature of quality. Also, unlocking requires deciphering all the tangible indicators and the interconnected human interactions, while interpreting them within a context.

Figure 3 shows the process in which those stakeholders interact and work toward programme quality. In the diagram, the teaching/learning environment where trainers and trainees interact will be the main focus of the investigation. Beyond that, the way the external factors and context variables affect learning and teaching are also examined.

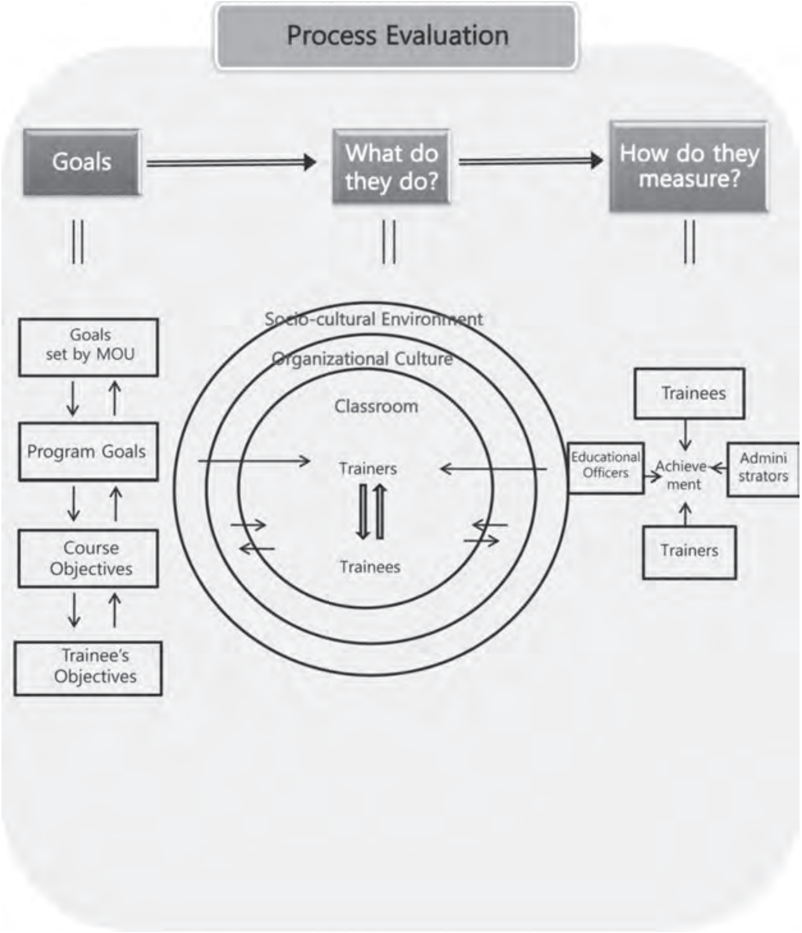


Figure 3 - Process Evaluation (Yeum 2011).

4. Conclusion

The paper presents a model to evaluate any INSET programme in similar EFL settings. First of all, the way different stakeholders set priorities for the particular INSET curriculum need to be examined to see if they share common objectives. Teacher trainees' needs should be triangulated with the perceptions and understanding of other parties. When their perceived needs are found to be in common, there is a higher possibility for the INSET programme to set up an appropriate curriculum.

Secondly, the way in which teacher trainers' instructional practices reflect what they believe can be found out through classroom observation. At the same time, what the teacher trainers do and why they do it needs to be reviewed from the perspectives of the trainees, who are teachers themselves. The teacher trainees can provide judging criteria to evaluate the trainers' instructional practices. Their own teaching beliefs and practices could be different from the trainers. The extent of agreement between the two parties' perceptions of goals and intentions will determine the overall value and learning outcomes.

In addition, the interaction modes between teachers and trainees with different socio-cultural and historical backgrounds should be investigated. The educational contexts in EFL settings may not be that socially, culturally, or historically challenging since trainers often deal with homogeneous groups of trainees. What is challenging within the particular teacher education setting, for example, could be that the teacher trainers, all foreign faculty staff, have to prepare Korean teachers to improve their teaching practice in their classrooms. In fact, the Korean trainees, experienced teachers themselves, may better understand the context variables in their classroom settings. In that case, the kind of strategies both parties develop to achieve the programme goals can be one of the elements to review in the programme.

Thirdly, how the success of the programme is measured by the different stakeholders needs to be investigated. The way each stakeholder measures programme quality and their perceptions of quality should be examined. The value and reliability of each method has to be triangulated among educational officers, programme administrators, and trainers and trainees. The way they measure success reflects each party's beliefs, and at the same time, can influence the instructional and administrative practices to influence programme quality.

Finally, cultural variables need to be discussed since they can affect the quality of the INSET programme. The general culture of the organisation should be examined to see how receptive the programme is to new ideas and suggestions. Any potential conflicts due to 'tri-cultural' aspects (administrators /teachers /learners) which impact quality should be also investigated. Potential culture clashes and misapprehensions between foreign trainers and Korean trainees, for example, might interfere with successful classroom interactions.

References

Crabbe, D. (2003). The Quality of language Learning Opportunities. ***TESOL Quarterly Vol. 37, No. 1, Spring, pp. 9-33.***

Crandall, J. (2000). Language Teacher Education. ***Annual Review of Applied Linguistics 20, pp. 34-55.***

Ellett, C. D. & Teddlie, C. (2003). Teacher Evaluation, Teacher Effectiveness and School Effectiveness: Perspectives form the USA. ***Journal of Personnel Evaluation in Education 17:1, pp. 101-128.***

Freeman, D., Orzulak, M., & Morrissey G., (2009). Assessment in Second Language.

Guskey T. R. (2002). Does it Make a Difference? Evaluating Professional Development. ***Educational Leadership, Vol. 59, No. 6. March, pp. 45-51.***

Hayes, D. (1996). "Prioritising "voice" over "vision.": reaffirming the centrality of the teacher in ESOL research. ***System, 24/2, pp. 173-86.***

Kiely, R & Rea-Dickins, P. (2005). ***Programme Evaluation in Language Education.*** Palgrave: Macmillan.

Long, M. (1984). Process and Product in ESL Program Evaluation. ***TESOL Quarterly, 18/3, pp. 409-425.***

Lynch, B. (1996). ***Language Programme Evaluation, Theory and Practice.*** Cambridge University Press.

Newmann, F.M., King, M.B., & Youngs, P. (2000). Professional development that addresses school capacity: Lessons from urban schools. ***American Journal of Education, 108, pp. 259-299.***

Pennington, M, & Young. A. (1989). Approaches to Faculty Evaluation for ESL. ***TESOL Quarterly, 23/4, pp. 619-646.***

Royse, D., Thyer, B., Padgett, D.K., & Logan, T. K. (2006). ***Programme Evaluation: An Introduction.*** Thomson: Brooks/Cole.

Tam, M. (2001). ‘Measuring quality and performance in higher education. ***Quality in Higher Education, 7(1), pp. 47-54.***

West, R. (1994). Needs analysis in language teaching. ***Language Teaching, 27, pp. 1-19.*** Cambridge University Press.

White R. (1998). What is quality in English Language Teacher Education? ***ELT Journal, volume 52/2 April 1, pp. 133-139.*** Oxford University Press. Oxford.

White R, Hockley A., Jansen J. & Laughner, M. (2008). ***From Teacher to Manager: Managing Language Teaching Organisations. Cambridge University Press. Cambridge.***

20

Research based approaches to assessment and evaluation: The English language teacher development programme

J.R.A. Williams and Rachel Bowden

Abstract

This paper outlines the background to the first phase of the Ministry of Education Malaysia/British Council English Language Teacher Development Programme. It discusses the aims, methods and problems arising from the research for the project baseline assessment. Learnings from this research are itemised and compared with some current literature. Observations are made concerning the impact on the project of existing assessment systems in primary schools. The programme approaches to teacher appraisal, and current and planned mechanisms for project evaluation are discussed. The paper concludes with a call for the empowerment and participation of stakeholders in assessment processes at all levels.

1. Background

The English Language Teacher Development Project (ELTDP), managed for the Malaysia Ministry of Education by the British Council, aims to improve the quality of teaching and learning of English in primary schools, increase teachers' English proficiency, identify and promote the use of materials to support the learning and assessment of English, and establish mechanisms to ensure project sustainability. This programme has been established to support the implementation of the new curriculum, the Kurikulum Standard Sekolah Rendah Bahasa Inggeris (KSSR), with its emphasis on formative assessment.

Essentially this is a mentoring programme with one British Council mentor working with 5 schools and focusing on the Level 1 teachers. 120 mentors will be placed in clusters totalling 600 schools throughout Sabah, Sarawak and the Federal Territory of Labuan for up to 3 years from 2011 to 2013.

2. ELTDP Phase 1

In the first Phase 1 for this project which encompasses the current study, 49 mentors established relationships with around 900 lower-primary teachers in 245 East Malaysian schools over a period of 10 weeks in order to conduct a participatory baseline assessment and needs and service audit. They were charged with:

- Informing stakeholders of the programme objectives
- Involving stakeholders in programme design through establishing the existing resources and services to support each objective, as well as perceived needs in each area
- Collecting information about schools, teachers, pupils, classrooms, parents
- Formulating their own impressions of the needs and services in their contexts
- Writing a report to summarise their findings.

2.1 Methods

Mentors used a wide range of methods in carrying out these research tasks: in-depth interviews, focus groups, and questionnaires with teachers, pupils, GBs, language officers, parents and community members, observations and document searches, as well as informal conversations within and beyond their schools. Researchers were asked to encourage teachers to help them survey and find out from children and parents what they thought their needs were, and what services (i.e. school, lessons, libraries, families and other English users, books, magazines, newspapers, TV, radio, computers) are available to equip pupils with the basic English language necessary to communicate effectively in a variety of contexts.

The process of data collection, situation and needs analysis had the positive effect of providing for the mentors an opportunity to understand the cultural differences between their backgrounds and those of the peoples of East Malaysia. They were able to see for themselves the constraints upon, and opportunities open to teachers, and develop an analysis of the specific local conditions of their schools and communities. It was intended that this process would lay the foundations to bring about change in schools, by encouraging teachers, GBs, parents and children to consider and assess many of the givens of school life and practice as a basis for validating or challenging practice. The mentors reported back to their informants with a summary of their conclusions and submitted to the project a detailed report on their findings.

The mentors were also invited to use their expertise and judgement to assess the teacher's needs in terms of language, teaching skills and competencies in order to translate the national curriculum into effective and appropriate classroom lessons.

2.2 Problems

It was at this point that flaws in the process became evident. The terms 'baseline assessment' and 'needs and services audit' suggest something objective. The intention was to make the process participatory and responsive which is why the project did not create any forms or specify instruments to be used. But when we asked mentors to make recommendations based on what they found, the frame of their previous experience dictated the way they investigated and reported. For example many mentors used a 'needs analysis questionnaire' of a type commonly found in the EFL classroom. These listed language and teaching competency areas and asked teachers to list which they wanted to focus on. The format dictated the answers and precluded the opportunity to explore what teachers understood by terms such as ‘classroom management’ or ‘grammar’, and how these connected to their classroom practice.

It became clear that that mentors were largely reporting what they thought, and had not sought means to ‘get inside’ the thinking of teachers, or that of the wider stakeholders. They relayed teachers’ observations on the barriers to effective teaching and learning without seeking to work with teachers to look beyond these. We saw that analysis of the curriculum was superficial at best, and mentors were uncritically accepting of teachers and others’ reasons why this or that could not be done. Researchers displayed little curiosity about the historical context of teacher development in the states and why previous teacher education had been ineffective in bringing innovation into the classroom, or about why the resources available (including libraries and ICT rooms) were so underused.

We were confronted with a new problem: if the research had been more participatory (and less reflecting of the views and values of mentors) would the results have been different? To test this we conducted a literature review and discovered that many of our conclusions were confirmed by previous studies.

3. Conclusions from Phase 1 with evidence derived from previous studies

The following conclusions were derived from the literature review:

- Many methods and approaches to primary teaching and ESL considered to be ‘best practice’ are contrary to values and beliefs held by teachers and other stakeholders. (Yaacob 2006, Pillay 2007)
- Some practices identified as negative, such as the lack of cooperation between teachers and collaboration between pupils, are also counter-cultural. (Hock & Raja 2002, Ghani 1992)
- Teachers do not express the rationale for their approach to teaching. Thinking around the nature of learning and language acquisition is concealed. Language is a barrier to some extent. (Pillay 2007, Yaacob 2006)
- Teacher perceptions of parental beliefs and attitudes vary from those expressed by the parents themselves. (Rajadurai 2010)
- No account is taken of the fact that many pupils in both Malay and Chinese medium schools have first languages other than these. (Smith 2003, Ting 2010)
- Schools vary in their resources, but whatever they have are generally underused. (Gardner & Yaacob 2009)
- There is a ‘standard’ English lesson which involves choral repetition and copying texts. (Yaacob 2006)
- Teachers report quite high job satisfaction (UNESCO-UIS 2008) but report that they do not enjoy teaching.
- There is a general assumption that children will have attended pre-school, but this is often not the case. (Masnan 2008)
- Whether a teacher is an ‘optionist’ or not does not predict their teaching capacity or level of language proficiency. (Ramli 2011)

4. Conclusions from Phase 1 not found in the literature

- There is no cooperation and coordination between subject areas in primary schools.
- There is a great deal of confusion about the new curriculum: teachers believing that they are prohibited from integrating skills; the primacy of listening and speaking; and especially misunderstanding the place of phonics in literacy development.
- Teachers' language proficiency is higher than their own self-assessment.
- "Co-curricular" and other non-teaching activities are often given priority over classroom teaching.
- Unattended classes are a common occurrence.

A final conclusion was drawn from the literature (Keshavarz & Baharudin 2009; Burns & Brady 1992; Bochner 1994) and has been applied to the plans for the project: that all prospective interventions should be approached with a reference to their cultural validity (Fleer 2006). The reader will understand that these cultural considerations were not foremost in the mind of mentors, most of whom derive their experience from an ELT context, which often assumes universally applicable teaching methods and activities. These will have been reinforced by the 'one-size fits all' approach of a national curriculum directed at the diversity of contexts and cultures offered by Malaysia.

5. Conclusions on the current role and impact of assessment in schools

A major finding of our reports provides evidence of the extraordinary impact testing has on the perceived purpose of schooling, on the allocation of resources, and what happens in classrooms. The project has noted that 'backwash' from the high stakes Year Six UPSR exam leads to pupils in both that and preceding years spending a great deal of time 'practising' for the exam. Ideas of formative assessment and assessment for learning do not permeate the classroom.

It is generally agreed that there is plenty of space for GBs to interpret the various ministry demands for assessment, and that they can be, and are 'very flexible'. So the message is that assessment (before the UPSR exam) is entirely in the hands of schools, and the idea is that teachers will create their own assessment instruments.

In practice assessment is largely a replication of the UPSR format of multiple choice and closed comprehension questions transferred downwards to the beginning of Year One. The new curriculum echoes that which it replaces in promoting "multiple sources of evidence like checklists, observations, presentations" and lessening the dependence on testing. But in many schools, regular pen and paper tests on a monthly and sometimes weekly basis punctuate the programme of teaching. While Ministry officials (from outside the Assessment Division) bemoan the fact that the "UPSR is killing our system", and predict that this venerated national institution will be abandoned, one school, not untypically, greets visitors with a board counting down "163 days to the UPSR"; mothers compare notes on their children's prospects for success in the exam years ahead; and children are commonly introduced by their UPSR grading rather than their name.

Reforming assessment requires deeper changes within the education system and will demand far more that abolition of state mandated exams. The process of change must be the result of an in depth partnership between pupils, families, parents, head teachers and especially teachers so that deeply embedded habits and routines are identified, questioned and compared with what people actually want for their children in education. UPSR may indeed be “killing our system”, but it remains the system for the majority of people concerned with primary schools, and until it is replaced, it will continue to determine much of what happens in classrooms.

The ELTDP plans to address this problem through intensive work with teachers to help them challenge the assumptions underlying their current attitudes to assessment and evaluation. We hope to be able to construct with teachers, and communicate clearly to pupils and parents, an understanding and application of processes of assessment for learning. Through these means we aim to make continuous assessment the most effective tool in forming decisions about future directions and activities, and in doing so bring renewed energy and vigour to the system.

6. Programme approaches to assessment: Evaluation of teachers’ progress

As a contextualised example of formative and continuous assessment based around multiple sources of evidence, the ELTDP has adapted a teacher Record of Achievement (RoA) which is designed to encourage teachers to develop habits of reflection and self-evaluation.

The RoA comprises one page each for preparation, implementation and evaluation. Each page is divided into individual criteria, and a statement labelled 0 which describes the expectations of an individual who has had no guidance in that area of teaching. Four statements follow this which identify stages of development, each stage subsuming the preceding statements. These statements are purposely very general, and will require further interpretation to account for the individuals involved in the procedure, the particular school circumstances, and other factors in the specific context of the record.

The idea is that partners (the mentor and the teacher, and possibly colleagues, head teachers, Language Officers etc.) will be involved in a consultation process leading to an agreed RoA. The mentor works with each teacher to complete the profile by working through the statements from left to right, marking each with a tick, a cross or a question mark, and recording evidence. Statements that cause problems or disagreements between the partners are discussed and areas of strength and weakness identified. Statements for particular attention in the short and longer term are decided upon.

This process is repeated at least twice yearly providing an agreed profile of the teacher’s progress and present stage of development. The profile is never used for comparing teachers, there is no such thing as a “Stage 2” or “Stage 3” teacher. Each teacher profile is unique, just as each teacher’s experience is unique. There is no expectation that a teacher will move seamlessly from one stage to the next as there may be new and different challenges. It may even be that a teacher moves down a stage due to differences in context, class size, school expectations or professional support.

In these aspects, the Record of Achievement satisfies many of the criteria for a robust assessment system. It is criterion-referenced rather than normative; it allows subjects to 'pass' through stages when they are ready; its results are derived through consensus and agreement between the teacher, a trusted other and where possible a wider range of supporters. The outcomes of the assessment are the formation of a shorter and longer term plans for further development.

While the RoA concentrates on professional development, a second monitoring tool, Most Significant Change (Davies & Dart 2005), allows teachers to contribute more personal evidence of change and development. Essentially the process involves the collection of stories from teachers, and the selection of the most significant of these by teachers, mentors, managers and the project as a whole. It is one of the techniques adopted by ELTDP as a method for the participatory monitoring and evaluation of the project. But in asking teachers at regular intervals to look back over last the few months, and describe what they think was the most significant change in the quality of teaching in their school, it allows for reflection and self-assessment among teachers and provides case study data towards planning and constructing future activities.

7. Programme approaches to assessment: Evaluating the progress of the project

Most Significant Change is one of the tools identified to provide data on impact and outcomes that can be used to help assess the performance of the programme as a whole. MSC is participatory because many project stakeholders are involved both in deciding the sorts of change to be recorded and in analysing the data.

The project is faced with several barriers to utilising all the elements of theory-based impact evaluation (White 2009). The project cannot work in non-project schools to establish a control group. Establishing a robust causal chain will not be possible due to the purposely vague project objectives (to 'improve' teaching and learning, teachers language proficiency, the extent and use of resources), and the reasonable expectation of heterogeneity that the ELTDP will witness some form of output, outcome and impact. Measurement of attributable variation becomes difficult. However, in its Phase One the ELTDP has made a concerted effort to become acquainted with its context, and is committed to rigorous continuing factual analysis, and will design mixed methods to further effective, appropriate and sustainable approaches to monitoring and evaluating the project.

Early focus group meetings are involving stakeholders in answering the following question: 'How can we best assess the achievements of the programme objectives?' This generation of indicators will lead to a participatory impact monitoring system, and intermediate and final evaluations of the project. This will involve the investigator becoming directly involved in the activities they are evaluating. The team leader will bring external knowledge of approaches, other experiences of participatory evaluation, and also of other projects, problems and solutions that have been found. Stakeholders will be represented and involved in teams implementing the evaluation, and these investigators can explicitly report on their own answers to the questions they will be asking.

In this way Participatory Evaluation helps participants learn about parts of the project in which they are not directly involved. Participatory Evaluation activities are an opportunity to build the capacity of stakeholders to increase participation in all activities and to equip them to sustain their involvement in similar activities after the project end. Recommendations from such an evaluation are more likely to be understood and implemented because those who will do the work were involved.

It is planned that the participatory evaluation process will begin with the identification and recruitment of a consultant who will be committed to several extended visits to the project starting at the end of the first year. Activities will include induction with mentors and ministry partners in order to reduce their anxieties and show the advantages of participatory evaluation, and to assist in planning the evaluation process. The team leader will then embark on a programme of participant orientation that will equip project workers and beneficiaries with tools and methods with which to engage the broadest possible constituency in contributing to the project evaluation.

8. Conclusions

The pursuit of objectivity in assessment posits an unachievable goal at the expense of denying agency to the people concerned. Improvement in teaching and learning depends ultimately on the stakeholders described in this paper, and it is inappropriate that they should be excluded from the process of researching their situation and measuring their progress. We believe that by empowering teachers to evaluate current practice towards identifying areas for development and establish indicators for measuring change, we will best utilise the connection between assessment and learning.

With the support of the Ministry of Education, the ELTDP has been offered the opportunity to realise this and has begun to invest in the principles of participation, relationships and sustainability which will remain the core of its work and the measure of its success.

References

Burns, D. & Brady, J. (1992). Cross-cultural comparison of the need for uniqueness in Malaysia and the United States. ***Journal of Social Psychology, 132, pp. 487-495.***

Bochner, S. (1994). Cross-cultural differences in the self-concept: A test of Hofstede's individualism/ collectivism distinction. ***Journal of Cross-Cultural Psychology, 25, pp. 273-83.***

Bolhasan, R. (2009). A Study Of Dyslexia Among Primary School Students In Sarawak, Malaysia ***School of Doctoral Studies (European Union) Journal 1.***

Davies, R. and Dart, J. (2005). ***The 'Most Significant Change' (MSC) Technique: A Guide to Its Use.*** http://www.mande.co.uk/docs/MSCGuide.pdf

Fleer, M. (2006). 'The cultural construction of child development: creating institutional and cultural intersubjectivity'. ***International Journal of Early Years Education 14: 2, pp. 27-140.***

Gaies S & R. Bowers (1993). Clinical supervision of language teaching: The supervisor as trainer and educator. In Richards, J. & Nunan, D. (Eds) ***Second Language Teacher Education.*** Cambridge: Cambridge University Press.

Gardner, S. and Yacoob, A. (2009). 'CD-ROM multimodal affordances: classroom interaction perspectives in the Malaysian English literacy hour'. ***Language and Education, 23: 5, pp. 409-424.***

Ghani, G (1992). Sustaining Curriculum Innovations Through Contextual Change. Paper presented at the Sixth Annual Conference of the Educational Research Association, Singapore (quoted by Pillay 2007).

Hock G. and M. Raja (2002) 'Exploring the Potential of Collaboration on Cooperative Learning among Educators from Different Institutions in Sarawak, Malaysia'. ***Asia Pacific Journal of Education, 22: 1, pp. 75-81.***

Hofstede, G. (1991). ***Cultures and Organizations***. Maidenhead: McGraw-Hill Europe.

Kling, Z. (1995). The Malay family: Beliefs and realities. ***Journal of Comparative Family Studies, 26 (1), pp. 43-67.***

Kon Yoon How (2008). 'Teaching Efficacy Beliefs of Primary School Teachers in Using English To Teach Mathematics And Science'. ***Jurnal IPBA. Jilid 3 Bilangan 2 45*** http://kajianberasaskansekolah.files.wordpress.com/2008/03/article4.pdf

Krishnan, U. (2004). ***Parent- adolescent conflict and adolescent functioning in a collectivist, ethnically heterogeneous culture: Malaysia.*** Doctoral dissertation, Ohio State University.

Masnan, A. (2007). ***Malaysian Preschool Education.*** Unpublished. Perak, Malaysia: Universiti Pendidikan Sultan Idris. http://www.scribd.com/doc/22301974/Malaysian-Preschool-Education

Mohamed, A., Lin Siew Eng & Ismail, A. (2010). Making Sense of Reading Scores with Reading Evaluation and Decoding System (READS) ***Canadian Center of Science and Education English Language Teaching Vol. 3, No. 3.*** www.ccsenet.org/journal/index.php/elt/article/download/7214/5565

Pillay, H. (2007). Working with Teachers in Challenging Teaching Contexts: Lessons Learnt in Powell-Davies, P. (Ed.) ***Primary Innovations Papers.*** Hanoi: British Council. http://www.britishcouncil.org/indonesia/indonesia-talking-english-primary-innovation.pdf

Ramli, A. (2011). Making an English out of science (ministry's policy). http://undomiel84.wordpress.com/2011/04/14/personal-professional-red-letters-part-ii-making-an-english-out-of-science-ministrys-policy/

Tafarodi, R. & Smith, A. (2001). Individualism–collectivism and depressive sensitivity to life events: the case of Malaysian sojourners ***International Journal of Intercultural Relations. Volume 25, Issue 1, pp. 73-88.***

Rajadurai, J. (2010). 'Speaking English and the Malay Community', ***Indonesia and the Malay World, 38:111, pp. 289 –301.***

UNESCO-UIS (2008). ***A View Inside Primary Schools: A World Education Indicators (WEI) cross-national study.*** Montreal: UNESCO.

Vethamani, E. (2008). ***Rigorous English Language Teacher Training Programmes Needed.*** http://edwinvethamani.com/blog/?p=13

White, H. (2009). ***What is impact evaluation, when and how should we use it, and how to go about it?*** Asia Development Bank/International Initiative for Impact Evaluation http://www.3ieimpact.org/userfiles/file/HWhite%20-%20
Introduction%20to%20IE%20-%20Dec%202009.pdf

Yaacob, A. (2006). ***Malaysian literacy practices in English: 'Big Books', CD-ROM and the Year 1 English Hour.*** PhD thesis, University of Warwick. http://wrap.
warwick.ac.uk/4076/1/WRAP_THESIS_Yaacob_2006.pdf

21

Demonstrating impact through cultural relations: How we evaluate British Council projects

Shannon West

Abstract

In 2009, the British Council adopted a logic model to provide greater structure to project evaluation and impact measurement. This paper describes the logic model and provides a case study to demonstrate its use through an educational change project in India that has been adopted nationwide as a result of a successful pilot evaluation.

Background

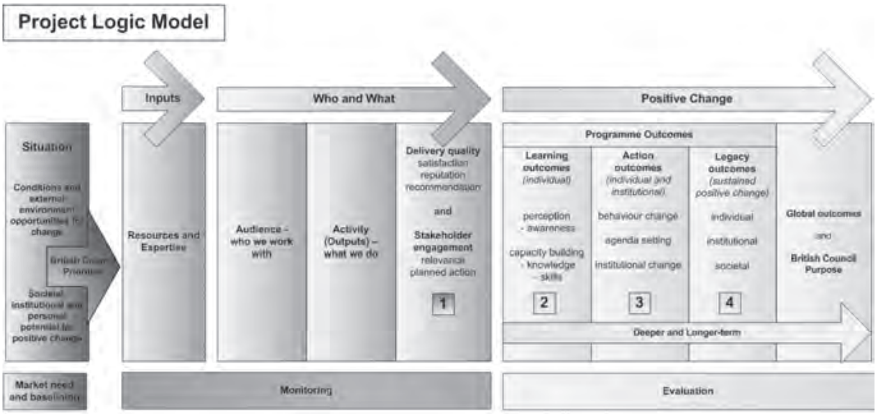
The British Council has managed a single project evaluation approach since 2002. In 2009 it refreshed the way it evaluates projects after reviewing good practice in other organisations. A central finding of this review was to establish a more comprehensive model to demonstrate results. This matches a growing trend both in the UK and internationally for more rigorous evidence of impact. A central output of this review was to introduce a project logic model.

The British Council logic model

A logic model lays out assumptions about how different elements of a project are linked together (see Figure 1) (Fretchling, 2007, Knowleton & Philips, 2009). It is a tool that can be used during project planning to describe how a project will work, and during project delivery to test planning assumptions and the degree to which project is on track. The British Council logic model shows the connections between:

- **situation** - the external environment, various stakeholder needs, our own strategy and how we decide what we would like to achieve
- **inputs** - what we invest in terms of people, time, money and partners
- **audience** – the people we work with
- **activity** - what activity we use to work with people

- **quality and engagement** – the immediate reaction and continued levels of engagement of the people we work with and
- **outcomes** - the change and impact that we plan to have with these people and their wider societies.



BRITISH COUNCIL © 2009 British Council, adapted from the ROI Institute™ methodology with assistance from abdi ltd, UK
Reference: Cooperative Extension of the University of Wisconsin-Extension (2003). Programme Development and Evaluation: Logic Model., Available: http://learningstore.uwex.edu/Planning-a-Program-Evaluation–P1033C0.aspx. Retrieved March 2009.

Figure 1 – The British Council Logic Model.

The logic model developed by the British Council is based on an educational model of change first developed by Kirkpatrick for evaluating the business impact of investing in the development and training of staff. This model was later extended to include a way to calculate return on investment (Phillips, 2007).

In order to institutionalise the approach, the British Council have worked to create the correct environment for evaluation (Stufflebeam, 2002):

- getting senior management buy-in by integrating the logic into organisational level performance indicators
- establishing the logic model in systems and processes, for example, project approval business cases
- communication and reporting against the logic model to demonstrate to staff the kind of information required
- training and staff professionalisation – the British Council has adapted the only UK-accredited evaluation qualification and trained over 70 staff.

How the logic model is used

When planning a project, we work from backwards starting with legacy outcomes (number 4 in Figure 1) or the sustained change we and stakeholders would like to achieve. For the types of projects that the British Council delivers this could be change for:

- **individuals** (for example, using the experience of a scholarship to achieve their potential in gaining employment/promotion in educational policy)

- **institutions** (for example, implementing a policy for managing carbon footprint in schools) and/or
- **societies** (for example, a greater tolerance of migration/migrants).

In order to achieve this sustained change we assume that the people we work with need to act or do something differently (number 3 in Figure 1). This could be a change in:

- behaviour of an individual (for example, putting their new educational policy development skills into practice) or
- an institution (for example, tabling the issue of carbon footprint management in meetings and in public, creating a policy for managing carbon footprint in schools).

In order to achieve the action outcomes above, we assume that the people we work with will need to learn or become aware of something (number 2 in figure 1). This learning change could be:

- a change in perception about the importance of something, or
- a new skill or improved knowledge (for example, improved knowledge of international educational policy, a changed opinion about the need for carbon footprint management, and so on).

In order for people to learn, we assume that we need to create a positive environment for learning and change to happen and continually check that we have the engagement of all stakeholders throughout the project (number 1 in figure 1).

Once we have our assumptions about how everything is linked, we can plan how to:

- monitor our progress along the chain of impact during delivery
- evaluate the degree to which we have achieved what we planned
- collect data about what else happened along the way, and
- assess what other influences there were over time that also may have contributed to any sustained change so that we can evaluate our contribution.

The logic model is a project level tool. However, as it sets out a uniform framework of categories of information, it can be used across a portfolio of projects and programmes and at organisational level to report impact. In order to achieve this, the British Council applies a number of principles. Firstly, although we may plan for sustained change, we may not achieve it in all contexts. The logic model can help stakeholders agree, with the levels of input and activity we can resource, what level of outcome can be achieved. Secondly, when reporting on deeper levels of impact, we must provide evidence of impact at all lower levels. Many British Council projects work in complex social environments over the long term. When sustained change is achieved, it is unlikely that the British Council and its stakeholders will have achieved it alone. Maintaining evidence along the chain of outcomes means we are able to link investment and activity to long term outcomes and attribute some of the resulting change. Finally, just because we plan for deep and sustained change, does not mean we evaluate to this level. Evaluation over the long term is time consuming and expensive. For the British Council delivering over 60 programmes in 110 countries, evaluating everything everywhere would be unworkable. The logic model, by outlining uniform categories of data, allows us to take a portfolio approach; evaluating to deeper levels in some locations only.

The logic model in application

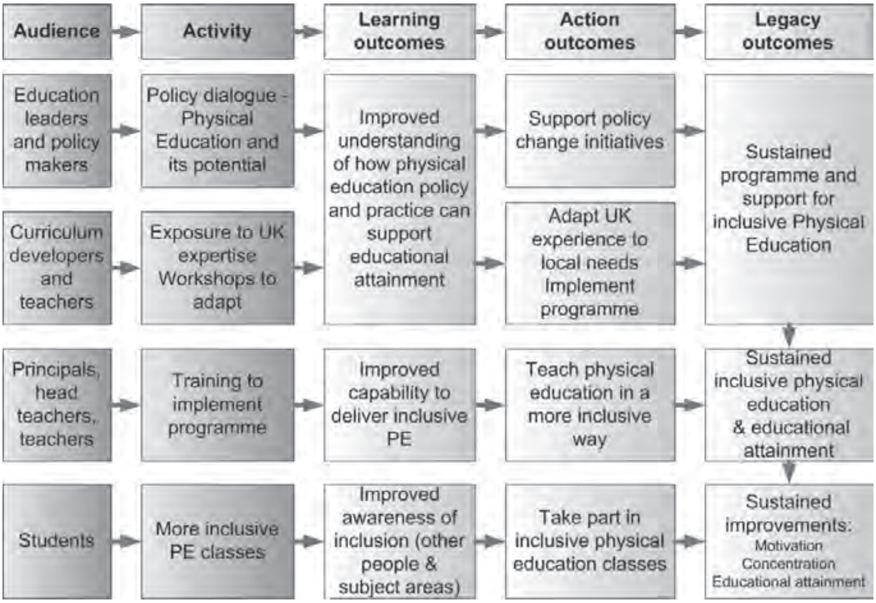
The following section describes a British Council project and how evaluation was used to gain support for wider, nationwide adoption of an educational change in India.

The *International Inspirations* project is the British Council's sports programme associated with the 2012 Olympic Games in London. Despite sports being an official part of school education in India, in practice, there was not a strong model to assist schools and teachers to implement good sports education. As a result, sport was not always included in the weekly syllabus. When sport was included, it was not always inclusive of students of differing physical and academic abilities and often left out girls and marginalised children.

The Physical Education Card project attempted to address the status of sport in Indian schools. A pilot project was initiated between the two national ministries that support school sports education (the Ministry of Human Resource Development, the Ministry of Youth Affairs and Sports) and various sports education and training organisations, the British Council and partners: UK Sports, UK Sports Trust and UNICEF. The UK had developed a model that was being used successfully to promote inclusivity in UK sports education. Practitioners from the UK worked with Indian counterparts to assess the suitability of the programme for India, make adaptations and plan a pilot programme to test its suitability in 58 Indian schools in 3 cities: Delhi, Mumbai and Chennai.

The programme was designed to work on multiple levels recognising the importance of different stakeholders: education policy makers, principals and teachers, students and parents. All of these people needed to be included and satisfied if the project was to be a success. The following outcome map (figure 2) based on the British Council logic model outlines the assumptions of how each stakeholder is involved and what they were expected to know and do. Ultimately, the goal was at the student level with sustained improvements in educational attainment and motivation through good sports education. However, the project needed to work with many stakeholders in order to achieve this goal. Setting outcomes for each stakeholder and monitoring and evaluating these was essential for success.

After gaining initial ministry level support for a pilot, the project team managed international collaboration between the UK and Indian education practitioners to design the pilot. At each stage, it was essential to make sure that overall outcomes of the project were understood and accepted and that people responsible for implementing the pilot on the ground were aware of their roles and what they needed to do to make the project a success. Interactive training workshops where teachers and principals experienced sports classes as they would be delivered to students were essential in bringing the project to life, demonstrating what was expected and gaining new skills and knowledge. At this stage, teachers were also introduced to project report forms that would be used to track implementation once the teachers returned to their schools. These evaluation report tools were essential in not only knowing how the project was going but also in making the project happen on the ground. Checklists and reporting requirements meant that teachers were continually aware of what was required of the pilot.



Audience	Activity	Learning outcomes	Action outcomes	Legacy outcomes
Education leaders and policy makers	Policy dialogue - Physical Education and its potential	Improved understanding of how physical education policy and practice can support educational attainment	Support policy change initiatives	Sustained programme and support for inclusive Physical Education
Curriculum developers and teachers	Exposure to UK expertise Workshops to adapt		Adapt UK experience to local needs Implement programme	
Principals, head teachers, teachers	Training to implement programme	Improved capability to deliver inclusive PE	Teach physical education in a more inclusive way	Sustained inclusive physical education & educational attainment
Students	More inclusive PE classes	Improved awareness of inclusion (other people & subject areas)	Take part in inclusive physical education classes	Sustained improvements: Motivation Concentration Educational attainment

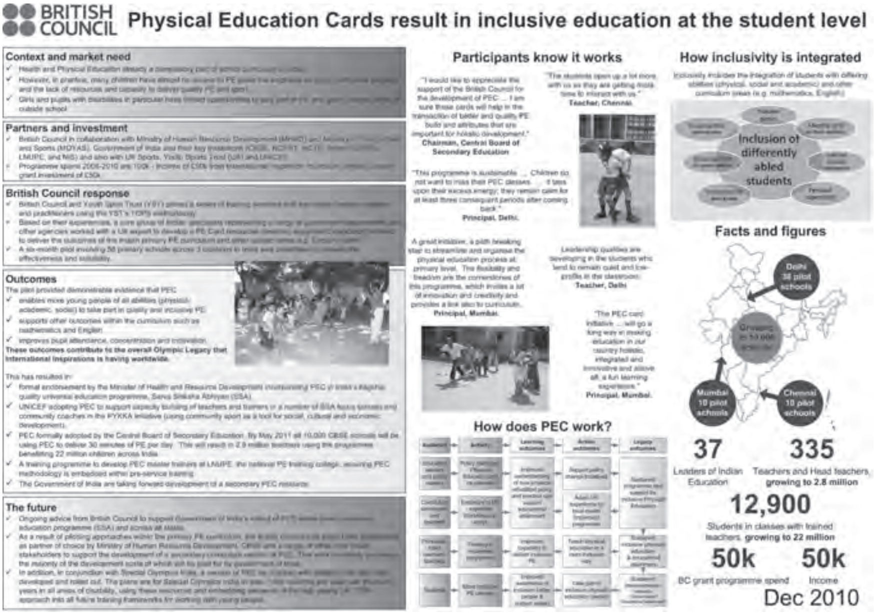
Figure 2 – Physical Education Cards outcome map.

As well as project reports from schools, independent Indian evaluation consultants were commissioned to visit all pilot schools to assess implementation, verify project reports from the schools and collect qualitative information about the programme from the students, parents, teachers and principals. Visits of senior policy makers to schools taking part in the pilot were also essential in maintaining engagement and demonstrating results first hand.

Within a short period of time, the pilot schools noticed impact among students including increased inclusivity of different students in sports activity, and increased motivation and concentration. Inclusivity covered students with differing physical and academic abilities. Some students who were less academically successful were recognised as leaders and given special tasks in sports activity by teachers and fellow students. The project evaluation highlighted areas for improvement if the project were to be rolled out nationally. It also identified critical conditions of a key stakeholder that had not been investigated - the lower status and pay of sports education teachers.

The pilot's success was partly due to the identification of the various stakeholder needs, having specific learning and action outcomes for these people and checking that they were being achieved throughout implementation. This allowed the project team to make assessments during implementation about whether the project was on track and make changes if not. Since the pilot concluded, the overall programme has been adapted using recommendations from the evaluation It has been officially endorsed and is being rolled out as part of the national curriculum.

The following evaluation summary report (Figure 3) was used within the British Council to highlight the achievements of the programme. It uses mixed media for example, diagrams, pictures, quotes and colour coding according to elements of the project logic model to highlight different types of information.



BRITISH COUNCIL **Physical Education Cards result in inclusive education at the student level**

Context and market need

- ✓ Health and Physical Education already a compulsory part of school curriculum [illegible]
- ✓ However, in practice, many children have almost no access to PE given the emphasis on [illegible] and the lack of resources and capacity to deliver quality PE and sport
- ✓ Girls and pupils with disabilities in particular have limited opportunities to take part in PE and sport [illegible] outside school

Partners and investment

- ✓ British Council in collaboration with Ministry of Human Resource Development (MHRD) and Ministry of [illegible] and Sports (MOYAS), Government of India and their key institutions (CBSE, NCERT, [illegible] LNUPE, and NIS) and also with UK Sports, Youth Sports Trust (UK) and UNICEF
- ✓ Programme spend 2006-2010 are 100k - Income of £50k from [illegible] grant investment of £50k

British Council response

- ✓ British Council and Youth Sport Trust (YST) [illegible] a series of training [illegible] and practitioners using the YST's TOPS methodology
- ✓ Based on their experiences, a core group of Indian specialists representing a range of [illegible] other agencies worked with a UK expert to develop a PE Card resource [illegible] to deliver the outcomes of the Indian primary PE curriculum and other subject areas e.g. [illegible]
- ✓ A six-month pilot involving 58 primary schools across 3 locations in India was [illegible] effectiveness and suitability.

Outcomes

The pilot provided demonstrable evidence that PEC

- ✓ enables more young people of all abilities (physical, academic, social) to take part in quality and inclusive PE
- ✓ supports other outcomes within the curriculum such as mathematics and English
- ✓ improves pupil attendance, concentration and motivation.

These outcomes contribute to the overall Olympic Legacy that International Inspirations is having worldwide.

This has resulted in:

- ✓ formal endorsement by the Minister of Health and Resource Development incorporating PEC in India's flagship quality universal education programme, Sarva Shiksha Abhiyan (SSA)
- ✓ UNICEF adopting PEC to support capacity building of teachers and trainers in a number of SSA focus [illegible] and community coaches in the PYKKA initiative (using community sport as a tool for social, cultural and economic development).
- ✓ PEC formally adopted by the Central Board of Secondary Education. By May 2011 all 10,000 CBSE schools will be using PEC to deliver 30 minutes of PE per day. This will result in 2.8 million teachers using the programme benefiting 22 million children across India
- ✓ A training programme to develop PEC master trainers at LNUPE, the national PE training college, ensuring PEC methodology is embedded within pre-service training
- ✓ The Government of India are taking forward development of a secondary PEC resource

The future

- ✓ Ongoing advice from British Council to support Government of India's rollout of PEC [illegible] education programme (SSA) and across all states
- ✓ As a result of piloting approaches within the primary PE curriculum, the British Council [illegible] as partner of choice by Ministry of Human Resource Development, CBSE and a range of other [illegible] stakeholders to support the development of a secondary curriculum version of PEC. [illegible] the majority of the development costs of which will be paid for by government of India.
- ✓ In addition, in conjunction with Special Olympics India, a version of PEC for children with [illegible] developed and rolled out. The plans are for Special Olympics India in [illegible] years in all areas of disability, using these resources and embedding [illegible] approach into all future training frameworks for working with young people.

Participants know it works

"I would like to appreciate the support of the British Council for the development of PEC … I am sure these cards will help in the transaction of better and quality PE build and attributes that are important for holistic development."
Chairman, Central Board of Secondary Education

"The students open up a lot more with us as they are getting more time to interact with us."
Teacher, Chennai

"This programme is sustainable … Children do not want to miss their PEC classes. … It takes upon their excess energy; they remain calm for at least three consequent periods after coming back."
Principal, Delhi.

A great initiative, a path breaking step to streamline and organise the physical education process at primary level. The flexibility and freedom are the cornerstones of this programme, which invites a lot of innovation and creativity and provides a link also to curriculum.
Principal, Mumbai.

Leadership qualities are developing in the students who tend to remain quiet and low-profile in the classroom.
Teacher, Delhi

"The PEC card initiative … will go a long way in making education in our country holistic, integrated and innovative and above all, a fun learning experience."
Principal, Mumbai.

How does PEC work?

How inclusivity is integrated

Inclusivity includes the integration of students with differing abilities (physical, social and academic) and other curriculum areas (e.g. mathematics, English)

Facts and figures

37 Leaders of Indian Education

335 Teachers and Head teachers **growing to 2.8 million**

12,900 Students in classes with trained teachers **growing to 22 million**

50k BC grant programme spend

50k Income

Dec 2010

Figure 3 – Physical Education Cards summary report to Executive Board.

Conclusion

Building the culture, systems and processes to support evaluation as a planning, design and implementation activity, (not just a post project activity), is challenging for many organisations. For the British Council the principal learning from the case study has been that evaluation (when integrated across the project planning, design and implementation phases and inclusive of specific outcomes for all stakeholders) provides a forum for a common understanding of what is trying to be achieved and of each stakeholder's specific role in making it happen. Evaluation also provides (especially when integrated with project implementation tools) the opportunity to keep track of whether we are likely to achieve the outcomes planned during delivery at a point when decisions can still be taken.

References

Fretchling, J. A., (2007). ***Logic Modelling methods in programme evaluation.*** Jossey-Bass, USA.

Kirkpatrick, D., (2011). Online. Available at: http://www.kirkpatrickpartners.com/OurPhilosophy/tabid/66/Default.aspx Retrieved 01 August 2011.

Knowleton, L. W. & Philips, C. C., (2009). ***The logic model guidebook.*** Sage, USA.

Philips, J. J., Philips, P. P., (2007) ***Show me the money: How to determine the ROI in people, projects, and programmes.*** Berrett Koehler, USA.

Stufflebeam, D., (2002) Online. Available at: http://www.wmich.edu/evalctr/archive_checklists/institutionalizingeval.pdf Retrieved 01 August 2011.

Contributors and Symposium Presenters

Abdul Halim Abdul Raof is Associate Professor at the Language Academy, University Teknologi Malaysia, Skudai, Johor. He holds a doctoral degree in Language Testing from the University of Reading and is currently Head of the test of English Communication Skills for Graduation Students Research Group of the university. He is a life member of the Qualitative Research Association of Malaysia (QRAM). He is a familiar face at international conferences having presented papers on language testing and evaluation. He supervises students on both undergraduate and post-graduate TESL programmes. His major areas of interest are in language testing, Academic Reading and writing, English for Specific Purposes and genre studies.

Abu Bakar Kaseh is a senior lecturer at the Department of Arabic Studies & Islamic Civilisation, Associate Fellow at the Institute of West Asian Studies and Deputy Director of International Relations Office of International Islamic University Malaysia. She holds an MEd from Pittsburgh and a PhD in educational management and evaluation, as well as other qualifications in language testing and evaluation. Her research interests include issues in language testing and assessment, the use of Modern Test Theory in test data investigation, and motivation in language learning.

Afrianto, from West Sumatra, went to Padang State University for his Bachelor degree in Education majoring in the English education programme, graduating with honours in 2000. He then taught for several years before he moved to Riau University in 2008. Apart from being an English teacher, Afrianto is also active as a freelance writer in some local and national newspapers in Indonesia. He is a member of the Curriculum Development Team of the English Study programme of Riau University.

Angela Hou Yung-Chi is Professor of Higher Education and serves as the director of the Development and Instructional Resources Centre of Fu Jen Catholic University as well as Dean of the Office of Research & Development of the Higher Education Evaluation & Accreditation Council of Taiwan (HEEACT). She specialises in higher education quality management, the internationalisation of higher education, faculty development, and quality assurance of cross border higher education. She has been conducting several national QA and ranking research projects for universities and the government over the past decade. Over the past 3 years, she has been in charge of international exchange affairs in HEEACT and engaged in many international activities of higher education quality assurance activities. She is also an Asia Pacific Quality Network (APQN) board member and consultant. She has published widely in the areas of higher education evaluation and rankings in local and international refereed journals.

Barry O'Sullivan is Professor of Applied Linguistics and Director of the Centre for Language Assessment Research (CLARe) at Roehampton University, London. He is particularly interested in issues related to performance testing, test validation and test-data management and analysis. He is very active in language testing around the world, working with government ministries and institutions, universities and examination boards. He has published widely on language testing and has presented his work at international conferences around the world. His books include 'Issues in Business English Testing' (CUP, 2006) and 'Modelling Oral Language Performance' (2008). 'Language Testing: Theories and Practices' (Palgrave) will appear in Spring 2011.

Clarence Jerry is an English Language lecturer of the Language Department at the Institute of Teacher Education, Tun Abdul Razak Campus, Kota Samarahan, Sarawak. He teaches Language Description, Academic English and ELT Methodology. His research interests are in the areas of EL studies/methodology and verbal protocol analysis.

David Carless is Associate Professor and Head of the Division of English Language Education in the Faculty of Education, University of Hong Kong. His main research interest is in how assessment can be reconfigured to stimulate productive student learning. Within this theme, Dr Carless carried out projects in Hong Kong schools, as well as in the tertiary sector. In relationship to assessment in higher education, in 2006 he co-authored (with Gordon Joughin, Ngar-Fun Liu and Associates) 'How assessment supports learning: learning-oriented assessment in action', published by Hong Kong University Press. With respect to schooling, his latest book is 'From testing to productive student learning: implementing formative assessment in Confucian-heritage settings', published by Routledge.

Douglas Sewell has been both a language teacher and teacher trainer for over 10 years in South Korea and China. Currently he teaches in the MA TESOL programme at Dankook University in Seoul, as well as being a programme tutor for the University of Birmingham's ODL programme. He also examines for a number of Cambridge exams including IELTS, and the TKT Practical Exam. Douglas' areas of interest are focussed on teacher training and language testing issues, as well as the process of self- regulation of language learning among ESL/EFL learners (central to his PhD studies through the University of Leeds).

Hyun-Ju Kim is an assistant professor of English Education at Dankook University in Korea, where she teaches undergraduate and graduates in TESL, language testing, and applied linguistics. She received her PhD in the programme of foreign Language and ESL Education at the University of Iowa. Her research interests are in World Englishes, L2 assessment, and the integration of the World Englishes perspective into the non-native speakers' English language proficiency tests.

J.R.A.Williams is currently a manager for the Ministry of Education/British Council English Language Teacher Development Project based in Sabah, East Malaysia. Previously he was a consultant for the Enabling Education Network (EENET), and education and early childhood advisor for Save the Children UK Middle East & North Africa, and then in Sri Lanka. He taught primary education at the University of the South Pacific and worked on English teacher development programmes in Oman, Hungary, Mozambique and Sudan. He holds a PGCE in TESOL, and an MEd in teacher training for ELT from the University of Exeter, UK.

Jim Tognolini is a Senior Research Fellow at the Oxford University Centre for Educational Assessment; Senior Vice President Research and Assessment for Pearson Global Strategies and Business Development; Professorial Fellow at Wollongong University (Australia); and Adjunct Professor of Education at the University of Western Australia. In his current position he is responsible for the design, implementation, and on-going management of assessment systems and learning process. He has advised and published widely on standards-based system of assessment.

Jin Kyung-Ae has been Senior Research Fellow and Director, Korean Institute for Curriculum and Evaluation since 1998. He holds an MA in linguistics and a PhD from the University of Pittsburgh.

John Field is a Senior Lecturer at CRELLA (Centre for Research in English Language Learning and Assessment) at the University of Bedfordshire, UK. He also teaches cognitive approaches to second language acquisition at the Faculty of Education, Cambridge University. He specialises in second language listening, on which he has written widely. His latest book, 'Listening in the Language Classroom' (CUP, 2008) won the international Ben Warren Prize for its contribution to language pedagogy. He brings a knowledge of psycholinguistics to much of his teaching and research and has written introductory books and a widely-used reference work 'Psycholinguistics: the Key Concepts' in this area.

John Hankinson is a British Council ELTDP Project Manager in Sarawak. His background is in primary teaching, and he has taught in the UK, Europe, Hong Kong and Brunei Darussalam, as well as in three UK universities. He has worked in primary level teacher development projects in Saudi Arabia, Brunei, Qatar and Sarawak, and has a particular interest in teacher mentoring and development as well as the use of modern resource banking techniques for teacher support.

Kashiraj Pandey is an Assistant Professor of English at Kathmandu University, Nepal. He has an MA in English Literature and MPhil in Education. He is an interpretive researcher in writing narratives and journaling as a means of transformation in teaching/learning. Kashiraj is the author of 'Writing Power'. For the last 12 years he has worked in teacher education and training for high school English Teachers in Nepal. He has presented several papers and given workshops in national and international conferences. He coordinates the training cell of NELTA.

Keith O'Hare has been working in the field of ELT for over 15 years. He has worked as a teacher and trainer in UK, France, Spain and China. He holds the Trinity TESOL Diploma. He has worked for the last three years in the Cultural and Education Section of the British Embassy Beijing, as English Projects Manager for national teacher training projects.

Kim Sang Jae is the Deputy Director of English Education Policy Division in the Ministry of Education, Science & Technology of the Republic of Korea. Dr Kim is currently in charge of supporting the development of the National English Ability Test.

Kyungsook Yeum has served as Administrative Professor responsible for several TESOL certificate programmes in Seoul, Korea. She has implemented a wide variety of language programmes over the past 14 years. In the process, she has gained deep understanding of the notion of quality control and a knowledge of how context variables influence programme quality. While serving as Korea TESOL's National President, she obtained a fuller understanding of the TESOL profession. Her latest experience is as the Conference Chair for PAC 2010 (Pan Asian TESOL Conference), "Advancing ELT in the Global Context". Currently, she is the Chair-elect for Program Administration Interest Group at the global TESOL Inc., and she is organising academic sessions on the theme "Quality Assurance in Language Teaching Organisations at Diverse Levels." She holds a PhD in English Literature and an MA in TESOL from the University of Maryland, Baltimore County. Currently, she is a PhD candidate in Applied Linguistics with the University of Macquarie.

Masnah Ali Muda is Deputy Director of the Examinations Syndicate at the Ministry of Education Malaysia. She holds a PhD in telematic technology and online learning. Dr Masnah has won numerous awards for her work in education and has presented at a variety of international conferences.

Masputeriah Hamzah is an Associate Professor at the Language Academy, Universiti Teknologi Malaysia (UTM). She is currently attached to the Office of Corporate Affairs of the University as the Deputy Director in charge of Corporate Communication and Branding. Her area of expertise is English for Specific Purposes (ESP). With more than twenty years of teaching experience, she now teaches ESP and Writing courses for the UTM TESL programme at both Masters and Undergraduate level. Her current research interests include certification of workplace written communication in English and establishing a profile for workplace oral communication. Other areas of interest include discourse analysis and syllabus and materials design.

Mina Patel has been active in education for twenty years. She is Managing Director of Ten Education Consultants Sdn Bhd, an English language training and consultancy company based in Malaysia. As well as the UK, Mina has worked in Thailand, Sri Lanka and Malaysia where she has worked as a Director of Studies, teacher and trainer trainer, and ELT projects manager. She holds a Masters degree in Applied Linguistics and Literature and her educational interests are in the areas of affect and motivation.

Pham Lan Anh is a lecturer in the Faculty of Foreign Languages at Hanoi Teacher Training College, Vietnam, where she teaches English Teaching Methodology, and Assessment and Testing. She is also a key trainer for Primary Innovations Project (part of Access English, run by British Council, Vietnam). Currently, she is doing PhD research on formative assessment. Her professional interests include teaching English to young children and teacher training.

Philip Powell-Davies is an independent education and social development consultant. He has 20 years' experience as a development educationist and social development expert with a research background in international education policy in developing countries, ELT, and the qualitative analysis of organisational change and capacity building, particularly in the public sector. He also has a strong interest in language policy and its links to education reform agendas. He has been a senior manager and country director of international organisations as well as holding senior positions as a ministerial advisor and programme director of national education development strategies. He is currently a senior advisory consultant with the World Bank, advising on education development in S Asia, and also leads an EU project examining the economic, political and social role and impact of global languages in the world today. He holds a PhD in international aid policy in education, as well as other degrees in social development management and education, and applied linguistics.

Samantha Grainger is currently Director English, British Council East Asia. She is responsible for the strategic management of British Council's English portfolio in East Asia and for managing the development and implementation of British Council strategy for English Business Development. Samantha has over fifteen years of teaching, training and project management experience in China, Egypt, Colombia, Poland and Japan. She holds a Masters Degree in Applied Linguistics from the University of Reading, UK.

Shannon West is Evaluation Manager for the British Council where he works on the use and development of the organisation's monitoring and evaluation system. He has worked for the British Council in English, education materials development, market research, and HR.

Shin Dongkwang, received his PhD in Applied Linguistics from Victoria University of Wellington, New Zealand. His expertise is in vocabulary learning and teaching, corpus linguistics, and language testing. He is working for the Division of English Education Research at Korean Institute Curriculum and Evaluation as a research fellow. He is in charge of NEAT (National English Ability Test) speaking and writing scoring and rater training.

Sophie Ioannou-Georgiou is currently an Inspector for English at the Cyprus Ministry of Education and Culture while also continuing to train teachers through the Cyprus Pedagogical Institute (in-service training) and the European University Cyprus (initial training). She is the co-author of 'Assessing Young Learners' (Oxford University Press) which was awarded the ELTON prize for innovation and excellence by the British Council and shortlisted for the Ben Warren Prize. She has extensive experience as an EFL teacher and teacher trainer. Her most recent publications are 'Guidelines for CLIL Implementation in Primary and Pre-primary Education' (co-edited with Pavlos Pavlou) and 'TESOL Technology Standards: Description, Implementation, Integration' (co-authored with Healey, Hanson-Smith, Hubbard, Kessler and Ware).

Sterling M. Plata is an assessment reform advocate who helps teachers, administrators and educational institutions to evaluate their assessment practices and to develop essential tools and processes that will deeply impact on student learning. She has a Specialist Certificate in Language Testing and Assessment from SEAMEO-RELC, Singapore. She is a full-time faculty member of the Department of English and Applied Linguistics, De La Salle University-Manila, and holds a PhD. Since 2000 she has trained more than 20,000 teachers and administrators all over the Philippines on assessment reform. She is the President of the Network of Language Teachers/Testers, Inc., and a member of the International Language Testing Association. Dr Plata is also an individual affiliate of the Association of Language Testers in Europe.

Tan Su Hwi is a Language Specialist with the Southeast Asian Ministers of Education Organisation (SEAMEO), Regional Language Centre (RELC). In this role, Dr Tan trains master teachers from different parts of Asia. She had also trained pre-service teachers in academic writing and communication skills at the National Institute of Education, Nanyang Technological University, Singapore. Her research interests include language policy and curriculum development, assessment, education philosophy as well as critical language discourse.

Vu Mai Trang (is Associate Dean, Faculty of English Teacher Education, University of Languages and International Studies, Vietnam National University, Ha Noi. Trang has extensive experience working as English teacher trainer and is co-developer of several teaching methodology materials for pre-service teachers at Vietnam National University and in-service teachers for Vietnam Ministry of Education and Training. She holds an MA in ELT from Nottingham University UK. She has published two translated books on literature, and her recent co-authored ELT book is ‘Learning English: A handbook for Primary Students and Their Parents ‘(Vietnam Education Publishing House, 2009).

If you are one of the many ministries of education or professionals grappling with the issues of how to effectively evaluate or assess English language learning, this collection of papers – the proceedings from the **British Council East Asia Regional Symposium on Assessment and Evaluation** – is essential reading. These Proceedings capture the insights from contributors with a wide range of backgrounds and experience, describing different contexts and tackling the issues from a variety of perspectives and as a result provide an immensely important contribution to discussions surrounding assessment and evaluation.

IELTS™

ISBN 978-0-86355-680-7

The British Council creates international opportunities for the people
of the UK and other countries and builds trust between them worldwide.