MerSETA Research Report

A preliminary investigation of the assessment model for Foundational Communication in English

Prepared for MerSETA by Melissa King

March 2010

EXECUTIVE SUMMARY

The concept of Foundational Learning has been set out in QCTO policy documents.

Foundational Learning (in the two learning areas of Foundational Communication in English and Foundational Mathematical Literacy) is intended to replace the current requirements for the Fundamentals, and is a compulsory achievement for final award of any occupational qualification from NQF 1 to 4.

The Foundational Learning pilot process during 2008-2009 piloted the implementation of learning programmes based on the Curriculum Frameworks in the two learning areas, and development of the assessment model. In order to meet the needs of stakeholders in relation to standardization, cost and human resource concerns, and an efficient and implementable assessment, a multiple choice assessment test format for the national external assessment of Foundational Learning was proposed.

The pilot process raised some concerns about the assessment model for Foundational Communication in English, in relation to perceived limitations of a multiple choice format for assessing English language skills.

This research report was commissioned in response to these concerns. It is seen as a preliminary investigation conducted by desktop research into the following areas:

- Correlations between text-based competence and listening and speaking skills.
- Correlations between reading and writing skills.
- International practice in relation to the multiple choice test format for the assessment of language competence.

The aim of the report is to give an overview of research in these areas, in a way that is accessible to the communities of practice who will be affected by the Foundational Learning project and who may not be multiple choice test specialists.

The study set out to determine whether there is sufficient evidence to support multiple choice assessment as an appropriate, realistic and useful model for the purposes of the Foundational Communication in English External Assessment. It summarizes features of multiple choice tests, samples correlation studies, and reviews local and international trends in relation to multiple choice assessments for English. In general, the findings in relation to correlations are positive, in the context of certain assumptions. These are primarily that (i) the items developed meet good practice criteria; (ii) the purpose of the FC external assessment is clearly understood, and candidates' results are used in appropriate ways; and (iii) there is ongoing research, monitoring and evaluation of the technical issues relating to this form of assessment , and of the rollout and impacts of the FC External Assessment, with a view to review and adjustment of the model.

The final section of the report discusses some of the findings in the context of the immediate needs and concerns of the Foundational Learning project.

CONTENTS

Part 1: Background

1. Foundational Learning: concept and background

The idea of Foundational Learning came about in the context of the development and design work done prior to the setting up of the Quality Council for Trades and Occupations[1] (QCTO). Much of this work focused on a new approach to the design of occupational qualifications and on the development of the Occupational Qualifications Framework (OQF) under the auspices of the Department of Labour (DoL).

Stakeholders in this process had ongoing concerns around notions of the kind of learning which underpins and supports occupational and technical training, especially in the areas of language competence (most especially in English as the medium of instruction for much occupational training in many sectors) and mathematical literacy. There had been much debate and some research on the role and impacts of the SAQA policy on compulsory Fundamentals, and it was felt that an alternative to this model was needed. The idea of Foundational Learning was developed[2], and ultimately described as follows:

> Foundational Learning refers to the competence identified as a platform for coping with the demands of occupational learning, in the two key areas of Communication and Mathematical Literacy. The aim of the Foundational Learning programmes is to deliver the identified skills to learners, while the aim of the Foundational Learning external assessment is to check whether learners have the competence described sufficiently in place. (DoL/ GTZ 2009:6)

Foundational Learning is intended to replace the current requirements for the Fundamentals, and is a compulsory achievement for final award of any occupational qualification from NQF 1 to 4. Foundational Learning is not located on a level on the NQF, as it is seen as 'foundational' to occupational learning. Any additional English or Mathematical requirements above Foundational are built into a particular occupational curriculum.

During 2007 a Curriculum Framework was developed in each of the two learning areas of Foundational Communication in English (FC) and Foundational Communication in Mathematical Literacy (FML).

2. The Foundational Learning assessment model

In order to support the needs of skills development in terms of cost, efficiency and time demands, the assessment model for Foundational Learning was conceptualised as follows:

[1] The QCTO was formally launched in February 2010.

[2] The process is described in the DoL/ GTZ Technical Task Team report of 2007.

- It would be a national, external, standardised assessment.
- Its purpose would be to provide a quick and efficient assessment to benchmark the broad competence level of an individual in the two Foundational Learning areas, in support of successful occupational training.
- It would be available whenever needed with a quick turn-around time for resulting.
- It would provide the basis for awarding certificates to learners in recognition of their successful achievement of Foundational Communication and/or Foundational Mathematical Literacy.

This model has subsequently been accepted as a working model by the QCTO.

In order to meet these requirements, it was proposed that the FC and FML External Assessments take the form of sets of items in a multiple choice format which are able to be machine-scored.

Other features of the model are discussed in Part 4, in relation to the research and findings of this report.

3. Impetus for this research: the Foundational Learning pilot

From April 2008 to March 2009, the Department of Labour and GTZ ran a pilot project on the Foundational Learning concept. This included implementation of Foundational Learning programmes based on the Curriculum Frameworks, and development of assessment items in both learning areas for trialling.

For Foundational Communication in English (FC) a matrix of assessable outcomes was identified from the FC curriculum. These were put into a Table of Specifications to guide the development of the items. This Table of Specifications and the structure of the FC External Assessment in its current form is attached as Appendix 1, and is still in the process of development and trialling.

The present research was prompted by comments made in the full research report (DoL/ GTZ Technical Task Team 2009) which was produced on both the curriculum and assessment findings of the pilot. In that report, it was recommended that research be undertaken on issues specific to the assessment of Foundational Communication (FC) in English through the medium of the multiple choice format.

The Pilot research report noted some limitations in relation to the language assessment:

- a written test will not assess the FC Element of Speaking and Listening; and
- while some writing outcomes can be assessed (through items such as pattern recognition, or selection of options which represent degrees of meaningful communication), clearly those outcomes which relate to the production of continuous prose cannot be assessed.

The report noted that while there were compelling arguments that the trade-offs represented by the FC Summative External Assessment could still allow it to achieve its stated purpose, some questions will only be answered in the context of ongoing development of the Assessment model, and long term research.

4. Purpose and scope of this report

It was felt that an initial response needed to be made to the concerns raised around the FC External Assessment in the Pilot report, in support of the rollout of Foundational Learning. While it was accepted that long term and in-depth research on what are complex technical and academic research questions needs to be built into future development processes around Foundational Learning, a limited literature survey on defined issues related to the FC assessment could be undertaken, and some expert opinion could be sampled.

On this basis the **research questions** were formulated as three areas for review:

- Correlations between text-based competence and listening and speaking skills.
- Correlations between reading and writing skills.
- International practice in relation to Multiple Choice Questions (hereafter referred to as MCQ) assessments for language competence.

The **methodology** followed was:

- Selective web-based international literature review
- Selective web-based South African literature review
- Informal telephonic or face-to-face discussions with a few selected individual experts[3].

The **aim** of this report is to serve as a resource for those who need to engage with issues around Fundamental Communication without necessarily being MCQ test or language experts. It is seen as a contribution to the development of common understandings for the communities of practice which need to actively participate in the rollout, monitoring and evaluation of Foundational Learning as an aspect of the occupational qualifications landscape. To this end, a brief description of multiple choice testing is included, and definitions of test terminology have been given as footnotes where appropriate.

It should be noted, however, that some constraints were experienced in the writing of this report. One problem was the highly scientific and technical nature of the discourse of relevant research. Obviously, the suitability and limitations of the MCQ test format for the assessment of language is an extensively researched area; however, MCQ research draws on its own disciplinary sources,

[3] With special thanks to Edward French for his guidance and input, and to Alan Cliff and Nan Yeld for their comments and source references.

such as item response theory and various statistical models for discussion of features such as reliability co-efficients, task variables and scoring distributions. The FC External Assessment development process must certainly be informed by long term research and tracking of this nature, and it is important that stakeholders respect this specialisation. It is equally important, however, that the implications and consequences of technical choices are mediated in understandable ways to those in the stakeholder community who will be supporting, implementing and benefiting from Foundational Learning, and who are not MCQ test specialists.

Another constraint is the context of much of the research into MCQ testing. Most correlation studies, for example, are located in schooling or in relation to the readiness of foreign or second language students for higher education in another language. Some of the implications of this are discussed throughout the report.

Part 2: Some general observations on MCQ testing

1. Summary of features

In order to give a context for subsequent discussion, a brief general description of MCQ tests is provided.

Multiple choice is a form of assessment in which respondents are asked to select the best possible answer from a number of choices. The questions are known as 'items'. Each item consists of stem and a set of options. The *stem* is the beginning part of the item that presents the item as a problem to be solved, a question asked of the respondent, or an incomplete statement to be completed, as well as any other relevant information. The stem can also include ancillary material such as case studies or visual stimuli or detailed descriptions. In the Foundational Communication (FC) external assessment, stimulus texts (for example, a reading text or a visual text) are linked to a set of items on that text.

The options are the possible answers that the examinee can choose from, with the correct answer called the *key* and the incorrect answers called *distractors*. Only one answer can be keyed as correct[4]. The following is given as an example.

EXAMPLE OF AN MCQ ITEM[5]

Feature	Example
Stem	The purpose of the last two paragraphs is to…
Options	
Distractor 1	▪ ensure that the reader knows how to identify stems.
Distractor 2	▪ introduce one to the idea of ancillary material such as case studies or visual stimuli.
Key	▪ describe the nature and general features of an MCQ item.
Distractor 3	▪ provide guidance into the choice of options in any MCQ test.

[4] Multiple response items can also be developed for more advanced tests.
[5] Thanks to Edward French

The parameters which define the quality of an item (and hence its re-usability) are discrimination and facility.[6] The quality of the test as a whole is measured by its validity (v) and reliability (r). It must be stressed, however, that high reliability does not mean high validity – a test developer could consistently be testing other objectives than those believed to be under scrutiny, a point we will return to later in discussion of the FC.

Item banking involves storing items and information about items (e.g. difficulty, discrimination, content codes, etc.) in electronic format. This allows searching based on parameters, and tracking item characteristics over multiple administrations[7]. The fact that statistics on computer-scored tests are easily obtainable is seen as an advantage of MCQ assessment.

Multiple-choice question tests are generally described as 'objective tests' in that there is complete marker objectivity in marking the test; objective tests are in general seen as having a greater reliability than open-ended assessments or essays. As pointed out above, this does not mean that objective tests are more 'valid' (for example, the items could consistently be testing irrelevant or nonsensical outcomes), but rather that this test format yields the same range of scores time and time again because there is a correct key answer.

It needs to be clearly understood as well that ' … the construction, specification and writing of the individual items or questions are influenced by the judgements of examiners as much as in any other test.'[8] Most discussions of multiple choice tests stress the fact that the quality of an objective test is determined by the skill of the constructors of the test. Item development teams need to include a number of high level skills and qualities: in the development of any items there are various forms of expertise at play, ranging from sound subject discipline knowledge and pedagogical imagination, through writing skills and professional/ technical/ scientific processing. Not least is an understanding of the target populations and the purposes and scope of the test.

Reliability increases with length of test and when questions with high discrimination values are included. Clear statements of instructions and items will also increase reliability.

Finally, there is the issue of the test construct and construct validity. There are many technical definitions and debates in relation to these terms. At the risk of oversimplifying, the construct refers to what one has set out to measure: the construct itself (on which the matrix or table of specifications used for designing the items is based) needs to be appropriate for the purposes of the test. A construct focusing only on spelling, for example, could not then claim to be assessing literary insight. In addition, there needs to be construct validity: the items developed for the test need to fit the construct put forward. For example, if all the items in a test deal only with literal

[6] Discrimination or D compares the number of correct responses to an item for the upper and lower performing segment of a group of test takers. Facility or F is the percentage of a group obtaining the correct answer, and this helps establish the acceptability of the level of difficulty of an item. This information informs storage for future use across different tests.

[7] Jay Parkes – www.flagugide.org/cat/multiplechoicetest. Accessed 29-01-10

[8] University of Technology Sydney www.iml.uts.edu.au/assessment/types. Accessed 25-01-10

comprehension of meaning against a construct which claims to assess critical language awareness, then there is a problem with construct validity. To put it in more technical terms:

> In psychometrics, 'construct validity' refers to whether a scale measures or correlates with the theorized construct it purports to measure. In essence it answers the question: "Are we actually measuring (are these means a valid form for measuring) what (the construct) we think we are measuring?" Http://en.wikipedia.org.

A summary of the advantages and disadvantages of MCQ tests follows[9]:

Advantages

- Multiple choice testing brings in economies and efficiencies of scale, because they are easily administered and quickly marked and resulted (utility, reliability and cost effectiveness)
- Well constructed items can tap into different levels of cognitive difficulty[10]
- Given the time efficiencies, multiple choice tests can sample a broad range of outcomes
- Scoring is objective and reliable
- Distribution of the scores is determined by the test, not by the examiner
- Items can be analysed and categorized, and stored in an item bank for future use
- Item tests yield useful statistical information.

Disadvantages

- MCQ items are time consuming and difficult to set, given all the variables around the quality of the stem and the distractors, and highly dependent on the item developer's expertise and experience
- There is no credit for partial information
- The tests may encourage guessing[11]
- There may be a negative washback effect on learning and teaching
- There are arguments around the limitations of the types of knowledge that can be assessed through this format.

[9] This section is summarized from various websites.

[10] Jay Parkes (www.flagugide.org/cat/multiplechoicetest) notes that 'Multiple choice items are flexible enough to tap into nearly any level of Bloom's taxonomy.' There does appear to be consensus that multiple choice items, if carefully developed in line with a viable construct, can access higher order skills.

[11] There are various ways of dealing with this in terms of item response theory and marking models which take this into account, and it is generally considered that the odds of a student receiving significant marks for guessing are very low when four or more options are available.

2. MCQ research context and implications for Foundational Communication external assessment

Much of the research done on MCQ tests has been undertaken in the context of large scale admissions test (extensively for ESL or EFL students) for high school equivalence, admission to college, or to university. This report draws mainly on such research. The implications of this for the FC External Assessment will be drawn out throughout the report.

The use of large scale admissions tests such as SATS or GMAT, IELTS or TOEFL[12] has been widespread internationally '… because it is believed that they will yield information about applicants' abilities to cope with the typical reading, writing and thinking demands they will likely face in Higher Education or that they will indicate the extent to which applicants will be able to cope with the language demands placed upon them in a particular medium-of-instruction.' (Cliff 2007:1-2).

In 1979 Cummins famously made the distinction between BICS (basic interpersonal communicative skills) and CALP (cognitive academic language proficiency). In spite of some controversy this distinction is still drawn on in both first language and ESL discourse today. BICS refers to the day-to-day language skills needed in social situations. This language is often determined by the context, it is not necessarily cognitively demanding and it is not specialised. Being proficient in BICS, however, does not necessarily mean that a student is proficient in using the same language in a learning context. The distinction is therefore made between BICS and CALP. The latter refers to using language for learning, which includes listening, speaking, reading, and writing about subject area content material. This is not just about understanding the content area vocabulary, but can include skills such as comparing, classifying, synthesizing, evaluating, and inferring. Information needs to be read, processed and presented in a more abstract context. Subject area information (whether this refers to general education areas or technical and occupational areas) is more cognitively demanding than the language used in social interactions. New ideas, concepts and language are presented to the students at the same time.

Cummins's distinction was originally made in the context of second language learners in schooling. In South Africa research around the concept of 'academic literacy' has most often been undertaken in the context of higher education[13]. The concept of academic literacy is linked to the understanding of the language as a vehicle for 'making meaning, making argument and understanding underlying points' (Cliff 2007:2). As Cliff notes, in international research the distinction has been drawn between a 'deep' and a 'surface' approach to using language for learning, in that 'deep' implies being able to access the underlying point or meaning of what is

[12] Scholastic Assessment Tests (SATs); Graduate Management Admission Test (GMAT); International English Language Test (IELTS); Test of English as a Foreign Language (TOEFL)

[13] However, current debate on the nature of vocational pedagogy implies the need for 'academic literacy' to become a theme in this context as well. See, for example, Gamble (2004) and Barnett (2006).

being read, while ‘surface’ implies taking in knowledge or facts as discrete and isolated pieces of information.

The fact that most of the MCQ testing discourse relates to contexts and levels which differ from that of FC context may mean that there are a number of different issues related to the construct and hence to construct validity. The question is, though, whether there is any common ground between these different contexts, so that we can feel comfortable making broad extrapolations from the correlation research to the Foundational Communication context.

The Foundational Communication (FC) curriculum certainly takes the notion of CALP (and, by implication, ‘deep’ learning) into consideration. It is stressed that the proficiency required for Foundational Communication, and any learning programmes developed against the curriculum, includes an emphasis on using English for learning, for understanding training materials and text books[14], and for being able to process and work with subject discipline content (whether these disciplines are the more general subjects needed in occupational training, or the technical subjects themselves). The goals of Foundational Learning are embedded in successful achievement in vocational and occupational training. The issue of vocational pedagogy is explored by both Gamble (2004) and Barnett (2006). Both writers stress the importance of transmission of formal conceptual knowledge supported by practical work. As Gamble notes:

> Practice needs to be related to theory in the same way as laboratory work is linked to the acquisition of scientific principles. Teachers have to be able to teach and instruct and students have to be able to render concepts in words, as is the case in all instruction concerned with the preparation of mind. It is not the non-languaged transmission through modelling that has long characterised traditional apprenticeship (2004 p 184-185)

So, although much of the literature found on ESL MCQ testing refers to academic literacy at a somewhat higher level for tertiary study, it would be acceptable to extrapolate some of the findings on language skill correlations at higher levels and/or different contexts to the FC assessment designed in the context of vocational and occupational training (always assuming that the MCQ items in this assessment themselves have construct validity against key features of the FC framework).

[14] For the development of the FC Curriculum Framework, a sample of NCV and industry-based training materials was reviewed in order to get an understanding of the language processing demands made on occupational learners. It was found that many text books and training materials in use were often written in a dense style, and require that readers be familiar with quite complex syntax and linguistic organizing devices in order to access the information.

Part 3: Trends in the use of MCQs for English Second Language (ESL) or English Foreign Language (EFL) MCQ assessments

ESL testing research covers a huge range of topics. This section takes a brief tour through some of the research done, specifically, on correlations between different language skills. Correlations refer to a reciprocal relationship between two or more things[15], and the degree to which they match. The discussion is held in the context of selected MCQ language tests, and refers to some of the current debates around these tests.

1. Correlations between different language skills

Research over the decades suggests that there are in general correlations on competence in the different skill areas. In relation to writing, Cooper makes an observation fairly typical of the ESL testing literature during the 80s and 90s:

> … although essay tests may sample a wider range of composition skills, the variance in essay test scores can reflect such irrelevant factors as speed and fluency under time pressure or even penmanship. When essay test scores are made more reliable through multiple assessments, or when statistical corrections for unreliability are applied, performance on multiple-choice and essay measures can correlate very highly …. at all levels of education and ability, there appears to be a close relationship between performance on multiple-choice and essay tests of writing ability. And yet each type of measure contributes unique information to the overall assessment. The best measures of writing ability have both essay and multiple-choice sections, but this design can be prohibitively expensive. (Cooper, 1984:1)

In the schooling contexts correlations between performance on MCQ items and performance on more extended production of text are evaluated via institutional evidence which is being gathered at the same time as the item testing evidence. In general, these correlations are found to be high: that is, those who perform well on the MCQ generally perform well on the writing production tasks, or as evaluated by teacher ratings or continuous assessment records.

[15] A correlation score is a statistical measurement of the relationship between two variables. Possible correlations range from +1 to –1. A zero correlation indicates that there is no relationship between the variables. A correlation of –1 indicates a perfect negative correlation, meaning that as one variable goes up, the other goes down. A correlation of +1 indicates a perfect positive correlation, meaning that both variables move in the same direction together. http://psychology.about.com.

As we shall see below, in the international literature there are calls for the re-introduction of direct assessment of the more active language skills, and explorations of cost-effective ways of marking these.

In the South African context some examples of MCQ tests are PTEEP (Placement Test in English for Educational Purposes), SATAP (Standardised Assessment Test for Access and Placement), and Test of Academic Literacy Levels (TALL). These are used in the context of HE for selection, placement and support needs identification.

The example explored in this report will be that of the PTEEP test. The University of Cape Town has undertaken research into alternative forms of admissions assessment over the past decade. The construct of the PTEEP test '…is conceptually constituted of nine sub-constructs that cover reasoning and meaning-making at a word, sentence, paragraph and argument level' (Cliff 2007:6), including a focus on visual and numerical literacy. The PTEEP test was not based on MCQ only, but included additional question types such as short pieces, an edit question and allowance for an essay. For the purposes of this discussion, the most interesting conclusions[16] drawn over the long term tracking of PTEEP was the finding that:

> The high correlations between various question types and the total score of writers on the test suggest that assessment using any one question type will suffice for determining the overall performance of writers. In particular, the multiple-choice questions on their own, or the short response pieces on their own, are very strongly correlated with the total score. (Cliff 2007:7)[17]

In sum, the test makers ' … have consistently found strong correlations between scores on multi-choice assessments and overall (total) scores of writers - this seems to imply that multi-choice assessment scores are 'capturing' what is captured in all forms of assessment combined (written; editing; short-response; etc.)[18]. Related research was done through UDW, in which '… the substantial numbers of students at UDW allowed us to run correlations between performance on MCQs and performance on other item types (short answers and essays). The correlations were consistently in the .9's, meaning that the two gave the same test information – one gained nothing in terms of a result by adding very expensive constructed responses'[19]. On the basis of this and related research, the newer benchmark test for UCT will be multiple choice only, given the economies of scale that this supports. These views give support to the notion that candidates' performance on the FC multiple choice will in fact be reflected to an acceptable degree in their production of writing.

[16] It is however also acknowledged that different question types do also yield a wider range of discrete information.

[17] Of relevance to this report and the FL focus on occupational and technical training is the fact that one of the studies done in relation to the PTEEP test was done with engineering students, where the findings were that performance on the PTEEP was in fact reflected by academic achievement in the course.

[18] Personal communication, Alan Cliff 2010.

[19] Personal communication, Nan Yeld 2010.

In international testing recent trends have shown a move away from total reliance on multiple choice testing to the inclusion of constructed responses. Brief discussions of the TOEFL test and the SAT will be used to illustrate these trends.

TOEFL stands for Test of English as a Foreign Language. The TOEFL was introduced in the 1960s by ETS (Educational Testing Services). It is probably the most widely used English proficiency test in the world. The TOEFL test measures English language proficiency in these three disciplines: reading, listening and writing. In most regions of the world TOEFL is taken on a computer (CBT: Computer-Based Test), though there is a paper-and-pencil version of the test in areas with limited access to computer-based testing facilities[20].

TOEFL and other similar tests relied for decades on traditional multiple choice tests of receptive skills. Although direct measures of speaking and writing abilities were incorporated into the Test of Spoken English (TSE) and Test of Written English (TWE) during the 1980s, these tests have always been used much less widely than the TOEFL test.

In more recent years there have been calls for a revision of the TOEFL test (most notably by the colleges and universities that use these for admission) to consider:

> … a new TOEFL test that: (1) is more reflective of communicative competence models; (2) includes more constructed-response tasks and direct measures of writing and speaking; (3) includes tasks that integrate the language modalities tested; and (4) provides more information than current TOEFL scores do about international students' ability to use English in an academic environment. (TOEFL 2000 (c):3).

The issue of washback[21] was also identified as an impetus for revision of the test:

> ESL/EFL teachers are concerned that discrete-point test items, and the exclusive use of traditional, multiple-choice items to assess receptive skills, have a negative impact on instruction. (TOEFL 2000(c):3)

[20] Since 2005 the Internet-based test (iBT) has progressively replaced both the CBT and paper-based tests in most countries (http://en.wikipedia.org/wiki/TOEFL).

[21] 'Washback' is the influence that a test can have on teaching and learning in the classroom, in that teachers teach to the test and learners learn what is to be tested. It can be either beneficial or harmful. Bailey (1999:3) notes: 'Definitions of washback are nearly as numerous as the people who write about it. These definitions range from simple and straightforward to very complex. Some take a narrow focus on teachers and learners in classroom settings, while others include reference to tests' influences on educational systems and even on society in general. Some descriptions stress intentionality while others refer to the apparently haphazard and often unpredictable nature of washback.'

This led to the TOEFL 2000 project which is aimed ultimately at the development of more specific frameworks and research agendas for the assessment of reading, writing, listening, and speaking, both as independent and integrated modalities.

The TOEFL Overall Framework paper lays out some of the issues that need to be addressed in such a revision process. For example, it notes that:

> Several Canadian, British, and Australian EFL tests do include some constructed-response tasks, but the annual testing volumes for these instruments are also quite low and the availability of testing centers and test administrations is relatively limited. Moreover, these tests have been criticized for their lack of evidence of scoring reliability and comparability of scores across test forms (Davies, 1987; Hamp-Lyons, 1987; Morrow, 1987; Rea, 1987; Tony, 1987; Alderson, Clapham, & Wall, 1995). Finally, there is little or no evidence that these tests are supported by an articulated theoretical model, or that they are more communicative or valid than the more traditional, multiple-choice tests. (TOEFL 2000 (c):4)

Two key issues in the historical favouring of multiple choice format are raised here: scale implementation constraints and issues relating to scoring reliability[22]. In addition, the report notes:

> The testing literature provides little guidance on how different response formats - multiple choice vs. constructed response, for example - actually affect examinees' performance on a test. Traub (1993) reviewed nine studies of the differences between multiple-choice and constructed response test items. Although some of the studies found differences between the formats, the differences were small, especially in the studies of reading comprehension. In summing up the research on reading tests, Traub concluded that "the answer is that [reading comprehension] tests that differ by format do *not* measure different characteristics" (1993, p.38, italics original). More importantly for construct validity, Traub found that none of the studies that found a difference could identify the construct differences between the two formats. (TOEFL 2000 (c):23)

The report goes on to note, however, that '… the greatest differences were found in studies that compared multiple-choice writing tests, such as the Test of Standard Written English (TSWE), with essay tests.' (TOEFL 2000 (c):23).

These issues are identified in the report as background to the project, not as reasons to suspend a revision of the scope and nature of the TOEFL test. What does become apparent in this report, though, is the extent of the research required to develop valid, reliable and efficient measures that can accommodate the new demands on the test. Different categories or subsets of research include, for example, the following: all the linguistic issues related to constructs, question types

[22] MCQ was favoured primarily because of the objectivity in the scoring, as opposed to subjective marking.

and task characteristics; scoring models; analysis of variables; research into the technologies that could support the new test, and could assist implementation of CBT. As the report notes, the many variables and complexities of the revision process meant that

> … it became apparent that a longer development and implementation time frame would be required to create, validate, and implement a new TOEFL test that responds to constituencies' concerns and needs. (TOEFL 2000 (c):6)

The 2000 Framework in fact provides the basis for a series of research monographs on frameworks for different skills sets and integration of modalities which collate research subsequently conducted up to 2006 (see references).

The issue of the scope and range of the research required for the development of the 2000 Framework and its sets of necessary components will be returned to in the section of the report that considers the immediate context and uses of the FC external assessment.

The SAT Reasoning Test is briefly described as another illustrative example in international testing. The SAT Reasoning Test (formerly Scholastic Aptitude Test and Scholastic Assessment Test) is a standardized test for college admissions in the United States. The SAT is owned, published, and developed by the College Board, a non-profit organization in the United States. It was formerly developed, published, and scored by the Educational Testing Service which still administers the exam. The College Board claims the test can assess a student's readiness for college. The test was first introduced in 1901, and its name and scoring have changed several times. The SAT Reasoning Test was introduced in 2005 and covers the three areas of maths, critical reading, and writing.

Of interest to this report is a brief description of the writing section. This includes multiple choice questions (worth about 70% of the score) and a brief essay. The essay was only implemented in 2005 following requests by colleges for examples of a student's writing ability.

The writing test is therefore weighted towards MCQ, and by implication claims that MCQ can assess certain facets of writing ability as follows:

> The multiple choice questions include error identification questions, sentence improvement questions, and paragraph improvement questions. Error identification and sentence improvement questions test the student's knowledge of grammar, presenting an awkward or grammatically incorrect sentence; in the error identification section, the student must locate the word producing the source of the error or indicate that the sentence has no error, while the sentence improvement section requires the student to select an acceptable fix to the awkward sentence. The paragraph improvement questions test the student's understanding of logical organization of ideas, presenting a poorly written student essay and asking a series of questions as to what changes might be made to best improve it. (http://en.wikipedia.org/wiki/SAT, accessed March 2010).

The assumption above is also made by the FC assessment model. The development plans of the agency currently piloting items for the FC External Assessment include the aim of introducing a wider range of question types than those trialled so far, in order to address the writing component in more depth.

The SATS essay section (as opposed to MCQ) is marked by two trained readers, and where there are score differences of more than one point a senior third reader decides. It should be noted that claims of bias have been made against the essay scores, once again raising the issue of scoring reliability in direct assessment of writing. Some of the criticisms have included the claim that length, and handwriting (where applicable,) impact on the score, and that 'general knowledge' shown in the essay affects the assessment of technical writing ability. In addition, a washback effect has been claimed:

> Perelman, along with the National Council of Teachers of English also criticized the 25-minute writing section of the test for damaging standards of writing teaching in the classroom. They say that writing teachers training their students for the SAT will not focus on revision, depth, accuracy, but will instead produce long, formulaic, and wordy pieces. "You're getting teachers to train students to be bad writers," concluded Perelman. (Perelman quoted in Winerip, 2005).

Issues of human resources[23] and scoring reliability are once again raised here, with the washback effect of open response seen (unusually) as negative.

The literature suggests that in general there are good correlations between performance in MCQ and ability to produce extended writing. But the historical trajectory in the two international tests discussed shows a movement away from reliance on MCQ only for language testing, in particular assessment of writing, and towards constructed and open ended tasks. It is, however, a movement not without its problems. In the South African context the trajectory has been very different, and we are only now re-discovering some of the benefits of MCQ for large scale testing. In the apartheid era South Africa had significant institutions dedicated to psychometric and edumetric testing. In the new era these were seen negatively and were downgraded (specifically in the educational community, though not perhaps in workplace assessment). Some recognition of their usefulness is re-emerging[24].

The implications of these findings for FC external assessment, looked at in the context of our needs, aims and constraints, will be considered in Part 4.

[23] ETS has developed a system known as e-rater®, the automated essay scoring for the marking of TOEFL. See Chodorow and Burstein 200, 3 and Lee, Gentile and Kantor 2008.

[24] As evidenced by the UCT research.

2. Reading in relation to listening/speaking

The research on correlations between reading and oral skills is difficult to summarize for the purposes of this report, given the technical nature of much of the research, and the number of contextual variables amongst various target populations and their situations[25] that the research brings into play.

Two examples are briefly discussed. The first is the English Language Proficiency Test (ELPT). This tests reading and listening (writing and speaking are not assessed) through a multiple-choice examination based on texts and on responses to an audio tape. It is designed to assess the test taker's ability to use English in day-to-day interactions involving listening and reading, emphasizing functional, practical language. The target population is high school students who have lived in the US for at least two years and who have either come from a country whose primary language is not English or who comes from homes where English is not the principal language.

Research undertaken in 1995 found reasonable correlations between the reading scores and the listening scores.

> Based on data from the first operational administration of the ELPT in November of 1995, the coefficient alpha reliability of the Reading score was .91 and the reliability of the Listening score was .89. The correlation of the Reading and Listening scores was .83. Corrected for unreliability, the correlation of the two scores was .91. As expected, reading and listening skills are highly related but the corrected correlation is still less than 1.0. (Bridgeman and Harvey, 1999:4)

Also relevant are various studies on the TOEIC assessment. The Test of English for International Communication (TOEIC) was developed to measure the English language skills of non-native speakers in the international work environment. It tests listening comprehension through MCQ in response to a range of stimulus material on audiotape, and reading through MCQ on texts of varying length and difficulty[26].

The section scores for reading and listening were constructed in such a way that the two could be compared. In the first validity study (1982),

> '… the correlation between the sections was 0.769 for the analysis sample. This would indicate that each score provides somewhat different information about the examinee and justifies reporting separate scores. (1982: page 8)

[25] These are generally very different to the oral contexts related to the Foundational Learning target populations, and the goals of the Foundational Learning curriculum frameworks.

[26] Recently TOEIC has introduced internet-based Writing and Speaking tests.

At the same time, however, it suggests that the correlations between reading and listening performance are not meaningless. Subsequent research (Wilson 1993 and 1999) studied the relationship between TOEIC scores and performance on the Language Proficiency Interview (LPI). The LPI is a well-established direct assessment (interview-based) of oral language proficiency, developed by the Foreign Service Institute of the U.S. Department of State. The LPI ratings range from 0 (no proficiency) to 5 (proficiency equivalent to that of a well-educated native speaker). In both these studies scores on the TOEIC have been found to be strongly related to performance on the Language Proficiency Interview (LPI) procedure. Wilson concludes:

> Based on these findings, guidelines were developed for making actuarial[27] inferences from scores on the TOEIC test regarding probable levels of LPI-assessed oral English proficiency in representative testing contexts. The development and use of such guidelines, of course, does not indicate that the indirect and direct measures are "construct equivalent." However, statistically validated inferences about expected levels of performance on the direct measure can be drawn from knowledge of performance on the indirect measure, and vice versa. (Wilson 1999:2).

On the basis of these examples we can suggest that there will be some (though not perfect) match between performance on reading and listening, and in addition a reasonably good correlation between listening skills and speaking skills. (The possibilities suggested for FC by the listening test models are noted in Part 4.)

[27] Actuarial science is linked to probability theory; the implication here is that one can make broad judgements about a group which may not apply to a few individuals within that group. However, such inferences will in general apply to the majority of a target population.

Part 4: Putting it all into context: conclusions for the FC External Assessment

Assessment is in essence imperfect; in particular, a specific external assessment is by nature a sampling of abilities in an unreal context. This comment applies to real life observations of practice and to essay-type testing as much as it applies to objective testing. In addition, as Hughes pointed out, ‘Language abilities are not easy to measure; we cannot expect a level of accuracy comparable to those of measurements in the physical sciences.’ (Hughes,1989/2007:2). Having said this, any test constructor is ethically bound to ensure that there is sufficient fit between the test, its stated purpose and the uses to which the results will be put.

This concluding discussion will be somewhat tentative in nature, with the aim of red-flagging areas that will need close and ongoing monitoring. It must also be stressed that the whole project of Foundational Learning and the FC External Assessment is in its infancy; it arose in response to an immediate and pressing need identified by occupational learning stakeholders, and perhaps in this instance even imperfect action is better than a continuation of current dilemmas. The project must be viewed as developmental and situated within an unfolding process, with monitoring leading to corrective action. We are obviously not at the same point of development such as that suggested by the research agenda of the TOEFL 2000 Framework revision.

Areas for discussion at this point are, then, as follows:

- What is the context for the Foundational Learning assessment model, and the FC External Assessment in particular? Why was this model advocated, and how is it justified?
- What is the meaning and use of the generally acceptable correlations found in the research when related to this context?
- What might be the impacts of the model on teaching and learning, and on learners?
- What are the future areas of research or action that the QCTO and its partners might consider?

1. The goals and context of the FC External Assessment

The impetus for the notion of Foundational Learning and its assessment model has been briefly described in Part 1. The variability of levels of English language skills in trainee groups undertaking occupational training through the medium of English in the FET band impacts on the quality of training, and on the progress of individuals. In addition, the inclusion of the compulsory Fundamentals in occupational qualifications did not necessarily achieve the desired goals of ensuring learners had the required ‘fundamental’ competence in place; in some cases the requirements excluded individuals, and in others they were poorly implemented as they were not seen as relevant to occupational training.[28]. One of the key aims of the Foundational Learning project was to begin an intervention that would assist occupational training providers and their quality assurance bodies in dealing with such concerns, in a way that was practical and implementable – and in ways that address problems of scale, and that support common understandings of the issues involved in the interface between language and occupational training.

To this end the Foundational Learning curriculum frameworks in FC and FML were developed, as specifications for intensive and stand-alone learning programmes in the two subject areas for those who need bridging in order to cope successfully with occupational training.

Against this background, some form of assessment is required in order to:

(i) identify those at risk, so that they can be placed in the relevant Foundational Learning programme;
(ii) certificate those who do not need a Foundational Learning programme; and
(iii) certificate those who achieve FC or FML after undertaking a learning programme.

Obviously, the target populations for Foundational Learning are not in a homogenous learning environment – on entry to the initial assessment and, for some, a Foundational Learning programme, some of these learners will not be in any learning environment at all, others may be in a workplace (either in workplace training or not), still others may be in or have come from a range of private providers or public institutions. In the interests of practicability, and in the context of these varied learning settings, the same test will be used for all three of the goals noted in (i), (ii) and (iii) above. The contexts in which the Foundational Learning External Assessments can be taken are summarized in the table below.

Foundational Learning External Assessment (FC or FML)	**Foundational Learning Programme (FC or FML)**	**Foundational Learning External Assessment (FC or FML)**
Regularly available, undertaken whenever a learner is thinking of doing occupational training. Successful learners get statements of results which they carry through into their occupational qualification. [Learners who are obviously below ABET 3 will not cope with the FL assessments.]	Learners who are not successful in one or both of the FL assessments undertake the relevant FL learning programme/s either before or during occupational training.	After completing the relevant FL programme/s, learners do the relevant external assessment again.

[28] Issues around the Fundamentals, the conceptual development of the Foundational Learning project and further developments in the pilot process are described in the Department of Labour reports 2007 and 2009.

The purpose of the FC and FML External Assessments is then to check whether candidates have the general competence to cope with the demands of occupational learning, in the two key areas of Communication and Mathematical Literacy

As noted in Part 1, the MCQ model was chosen for reasons of cost, efficiency in terms of availability and turn-around time, and the need for a standardized measure. Given the costs, human resource implications and lack of comparability around alternative forms of language assessment in non-institutionalised settings, there is certainly a common-sense justification for this decision. Also, we should note the advantages of going for a model that is simpler to implement rather than making the mistake of going for a complex approach that will not be implemented.

What we need to interrogate, then, is whether some of the limitations of MCQ tests for language are acceptable for the context and purpose of the FC External Assessment – in short, will the test be sufficiently useful, or will there be unacceptably negative impacts (on individuals, or on the quality of teaching and learning) that undermine this usefulness?

2. Making sense of MCQ limitations and correlations

Firstly, the research seems to suggest that the limitations are not as severe as avid proponents of production performance testing might think. In the context of some of the principles of the original NQF, we have become imbued with notions of rich and alternative forms of assessment; while these have certainly enhanced our understanding of assessment and extended our practices, some of these were not always implementable for their required purposes. There seems to be agreement in the ESL literature that language items constructed in line with all the criteria for good practice test development can indeed assess high-level and varied skills, and need not be trivial.[29] In addition, some fairly substantial aspects of skills in writing can be tested directly through MCQ. It can also be argued that MCQ can better test additional aspects of language competence linked to recognition, insight and reasoning which a candidate may be capable of but may not be able to articulate.[30]

Secondly, the literature suggests that there are broad correlations between performance on one skill or modality and another, most particularly in the case of MCQ performance and production of writing. While performance on MCQ in relation to production writing skills must be accepted as ‘an approximate measure … not an accurate one’ (Hughes 1989/2007:3), proximation may well be acceptable for the purposes of the FC External Assessment. However, a note of caution needs to be introduced in relation to the notion of correlation: the studies reviewed are all located

[29] Although as we have seen the international community appears to be investigating ways of extending large scale testing beyond MCQ, or using MCQ in new ways.
[30] Personal communication, Edward French

in the specifics of a particular test and the needs of its target population, and have usually been supported by other local forms of evidence. In sum, correlations need to be understood in the context of a specific test, and in relation to other measures or observations. Most importantly, for any notion of correlation to be used, there has to be an assurance that the primary items in use are valid against the test construct itself – that is, they are assessing what they claim to be assessing, and what they claim to be assessing is worth assessing, In other words, the construct itself must be fit-for-purpose against the aims of the test, before any deductions can be made from correlation studies.

Another problem for the FC External Assessment is that it will be applied to a variety of target populations: while their commonality is located in the idea (primarily) of using English for progress in occupational training through the medium of English, their application of the language will differ quite widely in different occupational training contexts and workplaces. How this will affect future research in relation to correlations in the context of Foundational Learning will need to be considered.

3. Relationship between the assessment model and the curriculum

This brings us to another thorny issue in relation to the FC External Assessment, and that is the fit between the assessment and the FC curriculum framework. One way of looking at this is simply to define the FC External Assessment as a proficiency test, according to the following established definition:

> Proficiency tests measure a student's achievement in relation to a specific task he will later be required to perform. For example, does a student know enough English to follow a particular course given in the medium of English or to do a particular job requiring a use of English? Proficiency tests rarely take into account any syllabus which the student has followed, since they are concerned with future performance rather than past achievement and are often administered to students from various language learning backgrounds. (Heaton, 1976:xi)

When the FC External Assessment is initially administered to learners, it fits this definition neatly. However, when it is given to those who have completed a Foundational Learning curriculum, questions could be raised. While the assessment construct was derived from the FC curriculum framework, the latter is necessarily 'bigger' in terms of coverage than the FC assessment construct itself. We then need to ask:

- Does this mean that those who are successful in the initial FC assessment, and are therefore awarded the FC certificate without doing a course, do not have 'the same competence' as those who achieve it after doing a programme?
- Or could we say that, as a sampling mechanism, the FC External Assessment is sufficient for the different groups of learners' competence to be regarded as equivalent?

Assuming that the FC External Assessment is related to a valid construct and is made up of good quality items, and taking the existence of broad correlations with the active skills into account, we could probably answer yes to the latter question. This is however an area that probably requires further discussion.

Related to this is the issue of washback, and the concerns raised that the ‘limitations’ of MCQ will impact negatively on the quality of the FC learning programmes developed. Two things need to be stressed here: firstly, providers need to be educated in relation to MCQ, and shown that this format will tap into deep skills which learners will need to be prepared for; secondly, the FC Curriculum Framework includes programme-based assessment requirements which demand a range of evidence (including listening and speaking evidence) which may not be assessed through the MCQ. On the assumption that providers meet these requirements (and are monitored in doing so[31]), we can hope that the FC programmes will not become limited by ‘teaching to the test’.

4. Uses of test results

Finally, there is the very serious issue of the purposes to which test results will be put. If a test is flawed, and its results are used for decisions which have serious consequences for individuals, the question of test ethics arises. The FC External Assessment was not conceptualized as a ‘high stakes’ assessment. It was devised as a tool to help individuals, providers and stakeholders to address some of the problems around foundational skills in occupational training. It is not an access assessment: achieving Foundational Learning is not a pre-requisite to entry into an occupational training programme or learnership, and failing the initial assessment is an indication that the learner needs an additional support programme which can be done parallel to occupational training, and which will surely be of benefit[32]. However, it is clear that management decisions may come into play here, and any test results can be abused in particular contexts. It will be the role of the QCTO to ensure that neither the FC nor the FML External Assessments are tainted by inappropriate use of the results.

Concluding remarks

This section will conclude with an overview of the main points that can cautiously be drawn out of this research, reminding readers that the contextual parameters of the Foundational Learning intervention must always be borne in mind.

[31] This will come under the QCTO’s role of accreditation of providers for Foundational Learning.

[32] It needs to be remembered though that the achievement of Foundational Learning is a requirement for award of the qualification at the end of occupational training; also, if an individual does not succeed in the initial assessment, it is problematic to embark on occupational training unless simultaneously addressing FC or FML.

Firstly, it seems clear that MCQ testing can be justified as an acceptable approach to language testing for the purposes of Foundational Communication in English. But this statement must be grounded in a number of assumptions:

- The construct against which the items are developed must be fit-for-purpose, and must as far as possible meet the aims of the Foundational Communication curriculum framework.
- Item development must be done according to good practice criteria, so that nuanced and varied items are developed; MCQs must not be trivial and must address all components of the construct.
- MCQ tests cannot be developed by amateurs or without being thoroughly field tested in relation to a range of correlations.
- Foundational Communication programme delivery needs to be monitored to ensure that there is no negative washback effect.
- Uses to which FC external assessment results are put in various constituencies need to be monitored.

In the largest sense, these responsibilities all fall under the QCTO, which must ensure that its Assessment Quality Partner carries out its brief appropriately, and that implementation of both assessment and curriculum delivery follows the intentions of the model. In another sense, all QCTO partners need to take responsibility for the unrolling of Foundational Learning and its assessment.

Apart from the ongoing statistical tracking that is the primary responsibility of the Assessment Quality Partner, there are various areas of research in which SETAs and providers could participate, in partnership with the QCTO and its research agenda. A vital area would be to ensure that there is some kind of interaction between findings generated in the context of occupational training itself, and any review of the Foundational Communication curriculum outcomes or assessment construct, so that the experience of stakeholders informs further conceptual development to a greater degree. Some other examples of research areas are:

- How good is the match between the FC curriculum framework, the FC External Assessment and the language needs of learners in training in different industry sectors?
- How can we track and compare the performances (in language terms) of those who enter occupational training (i) having achieved the FC through the initial assessment only; and (ii) having gone through an FC learning programme? (This area of research could go some way to addressing the questions raised in 3 above.)
- Are there ways in which cost-effective listening tests could be developed, following some of the models indicated in this research? Could a generic audio test with multiple choice questions be prepared (national level), or would industry-specific listening assessments be more appropriate (developed and paid for in local contexts)?

The concluding comment to this report is that we should celebrate the introduction of a more streamlined and efficient approach to assessment for the purposes of Foundational Learning, while not losing sight of potential pitfalls and problem areas that we can address as this undertaking unfolds.

References and Bibliography

Bailey, Kathleen M. 1999. *Washback in Language Testing.* TOEFL Monograph Series MS-15 1999. ETS, New Jersey.

Barnett, M. 2006. ‘Vocational Knowledge and Vocational Pedagogy’. Chapter 8 in Michael Young, Jeanne Gamble (eds). *Knowledge, curriculum and qualifications for South African Further Education.* HSRC Press, Pretoria.

Bridgeman, B and Harvey, A. 1999. *Evidence related to the validity of the English language proficiency test.* ETS Research Report, February 1999. www.ets.org

Chodorow, M and Burstein, J. 2003. *Beyond Essay Length: Evaluating e-rater's performance on TOEFL essays.* ETS Research Report 73, February 2003. www.ets.org.

Cliff, A.F., Ramaboa, K. & Pearce, C. 2007. *The Assessment of Entry-level Students' Academic Literacy: Does it matter?* Ensovoort, 11 (2), 33-48.

Cooper, P L. 1984. *The assessment of writing ability: a review of research*. GRE Board Research Report GREB No. 82-15R. ETS Research Report 84-12. www.ets.org

Gamble, J. 2004. ‘A future curriculum mandate for Further Education and Training colleges: recognizing intermediate knowledge and skill’. Chapter 8 in Simon McGrath, Azeem Badroodien, Andre Kraak, Lorna Unwin (eds). *Shifting Understandings of Skills in South Africa*. HSRC Press, Pretoria.

Gergely Dávid. 2007. *Investigating the performance of alternative types of grammar items.* In Language Testing 2007 24 (1) 65–97. SAGE Publications UK.

Department of Labour/ GTZ Technical Task Team. August 2007. *Foundational Learning Certificate for Occupational Qualifications.* Work-in-progress document. Department of Labour supported by GTZ.

Department of Labour/ GTZ Technical Task Team: Hallendorff E, King M, Machard D, Oberholzer A. 2009. *Foundational Learning Pilot Project: Research Report*. Department of

Labour supported by GTZ.

Hamp-Lyons, L. 1990, 2nd edition 1997.'Second language writing: assessment issues' in Barbara Kroll (ed) *Second Language Writing: Research issues for the Classroom.* Cambridge, Cambridge University Press

Heaton, J B. 1975. *Writing English Language Tests.* Longman Handbooks for Language Teachers. London, Longman Group Limited.

Hughes, A. 1989, 2nd edition 2003. *Testing for language teachers.* Cambridge, Cambridge University Press.

Lee Yong-Won, Gentile Claudia and Kantor Robert. 2008. *Analytic Scoring of TOEFL® CBT Essays: Scores From Humans and E-rater®.* ETS Research Report RR-08-01, January 2008. www.ets.org.

Rupp Andre, Ferne Tracy and Choie Hyeran. 2006. *How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective.* In Language Testing 2006 23 (4) 441–474. SAGE Publications UK.

TOEFL Monograph Series. Cohen A and Upton T. 2006 (a). *Strategies in Responding to the New TOEFL Reading Tasks.* RR-06-06. Educational Testing Service, Princeton, New Jersey

TOEFL Monograph Series. Cumming A et al. 2006 (b). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL*. MS-30. Educational Testing Service, Princeton, New Jersey

TOEFL Monograph Series. Cumming A et al. 2000 (a). *TOEFL 2000 Writing Framework: A Working Paper*. RM-005. Educational Testing Service, Princeton, New Jersey

TOEFL Monograph Series. Enright M K et al. 2000 (b). *TOEFL 2000 Reading Framework: A Working Paper*. MS-17. Educational Testing Service, Princeton, New Jersey

TOEFL Monograph Series. Jones et al. 2000 (c). *TOEFL 2000 Framework: A Working Paper.* RM-00-3. Educational Testing Service, Princeton, New Jersey

University of Cape Town, Centre for Higher Education Development. 2008. Item Development Workshop for the IEB, 30-31 July 2008.

Wilson, Kenneth M. 1993. *Relating TOEIC scores to oral proficiency interview ratings.* TOEIC Number 1 Research Summaries. www.ets.org.

Wilson, Kenneth M. 1999. *Validating a Test Designed to Assess ESL Proficiency at Lower Developmental Levels.* Research Report ETS RR-99-23. www.ets.org.

Winerip, M. 2005. "SAT Essay Test Rewards Length and Ignores Errors". *The New York Times* (May 4, 2005). . http://www.nytimes.com/2005/05/04/education/04education.html

Wu Yi'an. 1998. *What do tests of listening comprehension test? – A retrospection study of EFL test-takers performing a multiple-choice task.* In Language Testing 1998 15 (1) 21–44. Sage Publications UK.

Yo In'nami and Rie Koizumi. 2009. *A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats.* In Language Testing 2009 26 (2) 219–244. Sage Publications UK.

APPENDIX 1: Foundational Communication in English: External Assessment

Please note: This model was developed during the pilot process in order to inform the first round of development and trialling of items. **The model is a work-in-progress and is under ongoing review.**

The **Table of Specifications** was derived from the FC Curriculum Framework and lists ten categories of outcomes to be sampled. This is the matrix against which each test paper as a collection of items is developed, representing assessment of the learner's ability to:

i. Identify main points
ii. Recognise supporting ideas and detail
iii. Make inferences
iv. Track connections between ideas
v. Understand structure and organisation of texts
vi. Understand information presented in a variety of visual forms
vii. Recognise different purposes and text types
viii. Understand language conventions and forms
ix. Demonstrate knowledge of writing conventions
x. Demonstrate knowledge of grammar and syntax

Items are classified at different **cognitive levels of difficulty** according to the following formula to ensure the appropriate spread across each paper.

Elementary	20%
Intermediate	60%
Advanced	20%

The **structure** for the test as follows:

Section	Content	Number of items
A	Extended reading text, maximum 600 words	25
B	Short texts, paragraphs or single sentences	20
C	Visual literacy tasks (e.g. flow charts, graphs, diagrams, advertisements, tables lists)	15
		Total of 60 items
		50% Pass Mark for Competence