1 Introduction

Purpose of the volume

Language testing in Europe is faced with increasing demands for accountability in respect of all examinations offered to the public. Examination boards are increasingly being required by their own governments and by European authorities to demonstrate that the language ability constructs they are attempting to measure are well grounded in the examinations they offer. Furthermore, examination boards in Europe are being encouraged to map their examinations on to the Common European Framework of Reference (CEFR) (Council of Europe 2001), although some reservations have been expressed within the testing community as to the comprehensiveness of this instrument for practical test development and comparability purposes.

Weir (2005a) argues that a more comprehensive, coherent and transparent form of the CEFR would better serve language testing. For example, the descriptor scales could take increased account of how variation in terms of contextual parameters (i.e. specific features of the Writing task or context) may affect test performance; differing contextual parameters can lead to the raising or lowering of the level of difficulty involved in carrying out the target writing activity represented by a Can Do statement, e.g. 'can write short, simple formulaic notes'. In addition, a test's cognitive validity, which is a function of the cognitive processing involved in carrying out a writing activity, must also be explicitly addressed by any specification on which a test is based. Without such contextual and cognitive-based validity parameters, i.e. a comprehensive definition of the construct to be tested, current attempts to use the CEFR as the basis for developing comparable test forms within and across languages and levels are weakened, and attempts to link separate assessments particularly through social moderation by expert judges hampered.

Weir feels that the CEFR is best seen as a heuristic device rather than a prescriptive one, which can be refined and developed by language testers to better meet their needs. For this particular constituency its current limitations mean that comparisons based on the illustrative scales alone might prove to be misleading given the insufficient attention paid in these scales to issues of validity. The CEFR as presently constituted is not designed to say

with any degree of precision or confidence whether or not tests are comparable, nor does it equip us to develop comparable tests. Instead, a more explicit test validation framework is required which better enables examination providers to furnish comprehensive evidence in support of any claims about the sound theoretical basis of their tests.

Examination boards and other institutions offering high-stakes tests need to demonstrate and share how they are seeking to meet the demands of validity in their tests and, more specifically, how they actually operationalise criterial distinctions between the tests they offer at different levels on the proficiency continuum. This volume represents a first attempt to articulate the Cambridge ESOL approach to assessment in the skill area of writing. The perceived benefits of a clearly articulated theoretical and practical position for assessing writing skills in the context of Cambridge ESOL tests are essentially twofold:

- Within Cambridge ESOL – it will deepen understanding of the current theoretical basis upon which Cambridge ESOL tests different levels of language proficiency across its range of test products, and will inform current and future test development projects in the light of this analysis. It will thereby enhance the development of equivalent test forms and tasks.
- Beyond Cambridge ESOL – it will communicate in the public domain the theoretical basis for the tests and provide a more clearly understood rationale for the way in which Cambridge ESOL operationalises this in its tests. It will provide a framework for others interested in validating their own examinations and thereby offer a more principled basis for comparison of language examinations across the proficiency range than is currently available.

We build on Cambridge ESOL's traditional approach to validating tests, namely the VRIP approach where the concern is with Validity (the conventional sources of validity evidence: construct, content, criterion), Reliability, Impact and Practicality. The work of Bachman (1990) and early work of Bachman and Palmer (1996) underpinned the adoption of the VRIP approach, as set out in Weir and Milanovic (2003), and it can be traced back to about 1993 in various Cambridge ESOL documents on validity.

We explore below how a socio-cognitive validity framework described in Weir's *Language Testing and Validation: An evidence-based approach* (2005b) might contribute to an enhanced validation framework for use with Cambridge ESOL examinations. Weir's approach covers much of the same ground as VRIP but it attempts to reconfigure validity to show how its constituent parts (context, cognitive processing and scoring) interact with each other. The construct is not just the underlying traits of communicative language ability but is the result of the constructed triangle of trait, context and

score (including its interpretation). The traditional 'trait-based' approach to assessment had to be reconciled with the traditional 'task-based' approach (the CUEFL/CCSE approach and to some extent traditional Cambridge approach). The approach adopted in this volume is therefore effectively an *interactionalist* position which sees the construct as residing in the interactions between the underlying cognitive ability and the context of use – hence the socio-cognitive model.

In addition it conceptualises the validation process in a *temporal frame* thereby identifying the various types of validity evidence that need to be collected at each stage in the test development, monitoring and evaluation cycle. A further difference of the socio-cognitive approach as against traditional approaches is that the construct is now defined more specifically. Within each constituent part of the validation framework, criterial individual parameters for distinguishing between adjacent proficiency levels are also identified.

The conceptualisation of test performance suggested by Weir (2005b) is represented graphically in Figure 1.1.

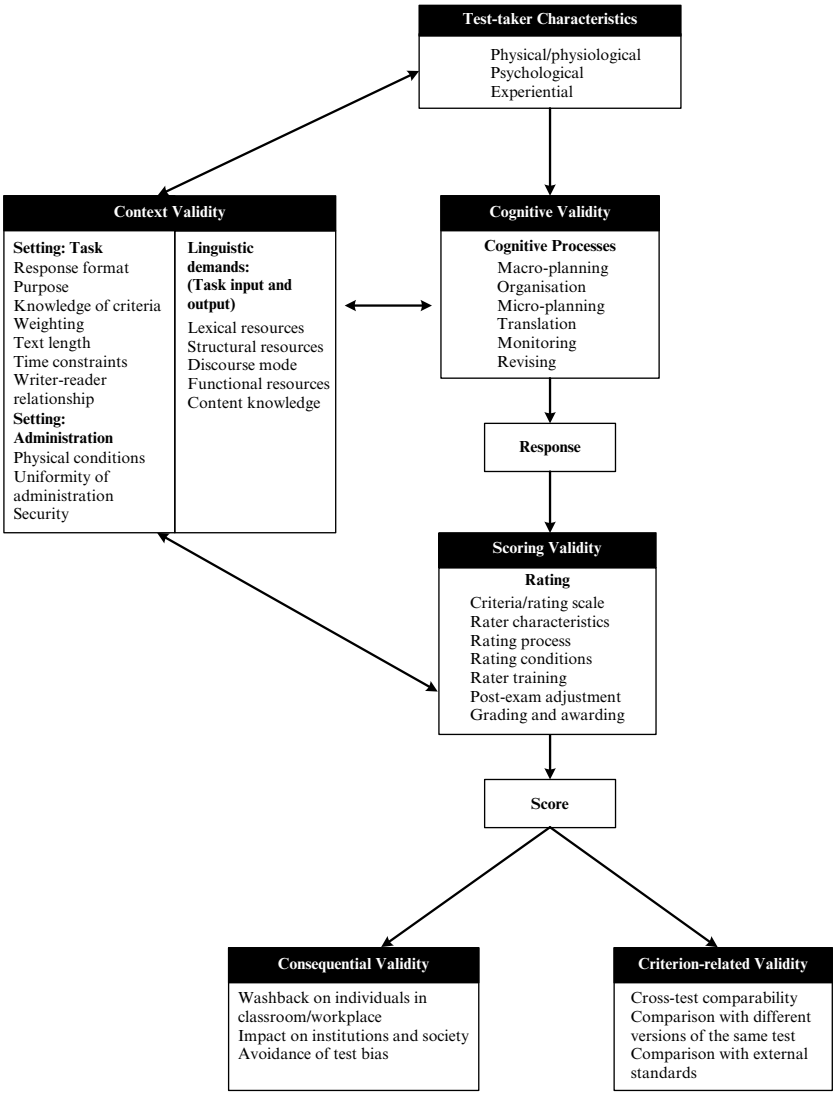
The framework is socio-cognitive in that the abilities to be tested are demonstrated by the mental processing of the candidate (the cognitive dimension); equally, the use of language in performing tasks is viewed as a social rather than a purely linguistic phenomenon. The framework represents a unified approach to establishing the overall validity of a test. The pictorial representation is intended to depict how the various validity components (the different types of validity evidence) fit together both temporally and conceptually. 'The arrows indicate the principal direction(s) of any hypothesised relationships: what has an effect on what, and the timeline runs from top to bottom: before the test is finalised, then administered and finally what happens after the test event' (2005b:43). Conceptualising validity in terms of temporal sequencing is of value as it offers a plan of what should be happening in relation to validation and when it should be happening.

The framework represented in Figure 1.1 comprises both *a priori* (before-the-test event) validation components of context and cognitive validity and *a posteriori* (after-the-test event) components of scoring validity, consequential validity and criterion-related validity. Weir notes:

> The more comprehensive the approach to validation, the more evidence collected on each of the components of this framework, the more secure we can be in our claims for the validity of a test. The higher the stakes of the test the stricter the demands we might make in respect of all of these (Weir 2005b:47).

A number of critical questions will be addressed in applying this socio-cognitive validation framework to Cambridge ESOL examinations across the proficiency spectrum:

Figure 1.1 A framework for conceptualising writing test performance (adapted from Weir 2005b:47)



- How are the physical/physiological, psychological and experiential characteristics of candidates catered for by this test? (focus on the test taker)
- Are the cognitive processes required to complete the test tasks appropriate? (focus on cognitive validity)

- Are the characteristics of the test tasks and their administration appropriate and fair to the candidates who are taking them? (focus on context validity)
- How far can we depend on the scores which result from the test? (focus on scoring validity)
- What effects do the test and test scores have on various stakeholders? (focus on consequential validity)
- What external evidence is there outside of the test scores themselves that the test is fair? (focus on criterion-related validity)

These are precisely the sorts of critical questions that anyone intending to take a particular test or to use scores from that test would be advised to ask of the test developers in order to be confident that the nature and quality of the test matches up to their requirements. The *test-taker characteristics* box in Figure 1.1 connects directly to the *cognitive* and *context validity* boxes because:

> these individual characteristics will directly impact on the way the individuals process the test task set up by the context validity box. Obviously, the tasks themselves will also be constructed with the overall test population and the target use situation clearly in mind as well as with concern for their [cognitive] validity (Weir 2005b:51).

Individual test-taker characteristics can be sub-divided into three main categories:

- physical/physiological characteristics – e.g. individuals may have special needs that must be accommodated, such as partial sightedness or dyslexia
- psychological characteristics – e.g. a test-taker's interest or motivation may affect the way a task is managed, or other factors such as preferred learning styles or personality type may have an influence on performance
- experiential characteristics – e.g. the degree of a test-taker's familiarity with a particular test may affect the way the task is managed.

All three types of characteristics have the potential to affect test performance.

The term content validity was traditionally used to refer to the content coverage of the task. *Context validity* is preferred here as the more inclusive superordinate which signals the need to consider not just linguistic content parameters, but also the social and cultural contexts in which the task is performed. Context validity for a Writing task thus addresses the particular performance conditions, the setting under which it is to be performed (such as

purpose of the task, time available, length, specified addressee, known marking criteria as well as the linguistic demands inherent in the successful performance of the task) together with the actual examination conditions resulting from the administrative setting (Weir 2005b:19).

Cognitive validity involves collecting both *a priori* evidence on the cognitive processing activated by the test task through piloting and trialling before the test event (e.g. through verbal reports from test takers), and also *a posteriori* evidence on constructs measured involving statistical analysis of scores following test administration. Weir stresses the importance of both:

> There is a need for validation at the *a priori* stage of test development. The more fully we are able to describe the construct we are attempting to measure at the *a priori* stage the more meaningful might be the statistical procedures contributing to construct validation that can subsequently be applied to the results of the test (Weir 2005b:18).

Language test constructors need to be aware of the established theory relating to the cognitive processing that underpins equivalent operations in real-life language use.

Scoring validity is linked directly to both context and cognitive validity and is employed as a superordinate term for all aspects of reliability (see Weir 2005b: chapter 9). Scoring validity accounts for the extent to which test scores are based on appropriate criteria, exhibit consensual agreement in their marking, are as free as possible from measurement error, stable over time, consistent in terms of their content sampling and engender confidence as reliable decision-making indicators.

Criterion-related validity is a predominantly quantitative and *a posteriori* concept, concerned with the extent to which test scores correlate with a suitable external criterion of performance with established properties (see Anastasi 1988:145; Messick 1989:16). A test is said to have criterion-related validity if a relationship can be demonstrated between test scores and some external criterion which is believed to be a measure of the same ability. Criterion-related validity sub-divides into two forms: concurrent and predictive. Concurrent validity seeks an external indicator that has a proven track record of measuring the ability being tested (Bachman 1990:248). It involves the comparison of the test scores with this other measure for the same candidates taken at roughly the same time as the test. This other measure may consist of scores from some other tests, or ratings of the candidate by teachers, subject specialists, or other informants (Alderson, Clapham and Wall 1995). Predictive validity entails the comparison of test scores with some other measure for the same candidates taken some time after the test has been given (Alderson et al 1995).

Messick (1989) argued the case for also considering *consequential validity* in judging the validity of a test. From this point of view it is necessary in validity studies to ascertain whether the social consequences of test interpretation support the intended testing purpose(s) and are consistent with other social values. There is also a concern here with the washback of the test on the learning and teaching that precedes it as well as with its impact on institutions and society more broadly. The further issue of test bias takes us back to the *test-taker characteristics* box. The evidence we collect on the test taker should be used to check that no unfair bias has occurred for individuals as a result of decisions taken earlier with regard to contextual features of the test.

Validity as a unitary concept

Although for descriptive purposes the various elements of the model in Figure 1.1 are presented as being independent of each other, there is undoubtedly a 'symbiotic' relationship that exists between context, cognitive and scoring validity, which together constitute what is frequently referred to as *construct validity*. Decisions taken with regard to parameters in terms of task context will impact on the processing that takes place in task completion. Likewise scoring criteria where made known to candidates in advance will similarly affect executive processing in task planning, and monitoring and revision. The scoring criteria in writing are an important part of the construct in addition to context and processing since they describe the level of performance that is required. Particularly at the upper levels of writing ability, it is the quality of the performance that enables distinctions to be made between levels (Hawkey and Barker 2004). The interactions between, and especially within, these different aspects of validity may well eventually offer further insights into a closer definition of different levels of task difficulty. For the purposes of the present volume, however, the separability of the various aspects of validity will be maintained since they offer the reader a helpful descriptive route through the socio-cognitive validation framework and, more importantly, a clear and systematic perspective on the literature which informs it.

Audience for the volume

This volume is aimed primarily at those working professionally in the field of language testing such as key personnel in examination agencies and those with an academic interest in language testing/examining. It is intended as a high level academic statement of the theoretical construct on which Cambridge examinations are based. As such it is hoped that it will offer other institutions a useful framework for reviewing their own examinations.

However, some parts of the volume may also be of interest and relevance to anyone who is directly involved in practical writing assessment activity and/or Cambridge ESOL examinations in some way, e.g. writing curriculum and materials developers, teachers preparing candidates for the Cambridge Writing tests, etc.

Voices in the volume

As the reader progresses through the volume, it will become apparent that there are several 'voices' in the book, along with various styles of expression.

First, there is the voice of the wider academic community in Applied Linguistics and Language Testing which provides the theoretical base for the framework we have adopted and the guiding principles on which we feel good practice should be based. In discussing each section of the above framework an account is first given of contemporary thinking on the parameter under discussion.

Then there is the voice of the language testing practitioners within Cambridge ESOL who are responsible for developing, administering and validating versions of the tests. Alongside this may be detected the voice of the large community of external professionals who are actively associated with the production and delivery of Cambridge ESOL tests (e.g. test item writers, Writing examiners, centre administrators, etc.).

These latter voices are referred to after we have addressed the current thinking on a particular element of the framework. Sometimes they take the form of case studies to exemplify particular issues, at others they exist in quotations from, or references to, external and internal documentation such as examination handbooks, item writer guidelines, examination and centre reports.

It will become clear that, in compiling the volume, we have drawn together important material from a variety of sources within the organisation relating to the operationalisation of Cambridge ESOL's exams in relation to the theoretical framework; some of this information is extracted from previously internal and confidential documentation and is appearing in the public domain for the first time. It reflects Cambridge ESOL's ongoing commitment to increasing transparency and accountability.

The presence of multiple voices, together with the assembly of information from a wide variety of different documentary sources, inevitably means that differing styles of expression can be detected in certain parts of the volume. Apparent shifts in voice or style simply testify to the complex network of stakeholders which exists in relation to any large-scale testing practice and the fact that any large-scale testing enterprise constitutes a complex, and sometimes sensitive, ecosystem (see Weir and Milanovic 2003 for further discussion of this).

Focus of the volume

Research into the assessment of second language writing normally concerns itself with the direct testing of language performance. By a 'direct test' we mean one which tests writing through involving candidates in the actual construction of text in contrast to 'indirect' or 'objective' tests of writing which principally focus on knowledge of microlinguistic elements of writing, e.g. through multiple choice, cloze, gap filling or error recognition response formats (Hyland 2002:8–9). In these indirect tests writing is divided into more specific 'discrete' elements, e.g. of grammar, vocabulary, spelling, punctuation and orthography, and attempts are made to test these formal features of text by the use of objective test formats. These tests are indirect in that they are only measuring parts of what we understand to be the construct of writing ability. What they test may be related to proficient writing as statistical studies have indicated (De Mauro 1992), but they cannot represent what proficient writers can do (Hamp-Lyons 1990). It would be difficult to generalise from these types of test to how candidates might perform on more productive tasks which required construction of a complete text. It would be difficult from these discrete item tests to make direct statements about how good a writer is or what he or she can do in writing.

As a general principle, it is here argued that language tests should, as far as is practicable, place the same requirements on test takers as are involved in writers' responses to communicative settings in non-test 'real-life' situations. This approach requires attention to both cognitive and social dimensions of communication. According to Hyland, the purpose for writing in this new paradigm is communication rather than accuracy. He argues that tasks within this paradigm are concerned with the psychological reality rather than statistical reliability (Hyland 2002:8, 230). Jacobs, Zinkgraf, Wormuth, Hartfiel and Hughey (1981:3) draw attention to the additional communicative dimension of writing as a social interaction with its emphasis on communicative purpose and the importance of the effect on the reader in the process. Hamp-Lyons and Kroll (1997:8) similarly emphasise that writing is a social and cultural act as well as a cognitive activity with *context, purpose* and *audience* as key parameters.

These views on direct Writing tasks (see Grabe and Kaplan 1996 and Hyland 2002 for excellent overviews of writing) reflect a concern with *authenticity* which has been a dominant theme in recent years for adherents of the *communicative testing* approach as they attempt to develop tests that approximate to the 'reality' of non-test language use (real-life performance) (see Hawkey 2004b, Morrow 1979, Weigle 2002, Weir 1983, 1993 and 2005b).The 'Real-Life' (RL) approach (Bachman 1990:41), though initially the subject of much criticism in the USA, has proved useful as a means of

guiding practical test development. It is particularly useful in situations in which the domain of language use is relatively homogeneous and identifiable (see O'Sullivan 2006 on the development of Cambridge Business English examinations).

With regard to Cambridge ESOL examinations, authenticity is considered to have two characteristics. First, *interactional authenticity*, which is a feature of the cognitive activities of the test taker in performing the test task (see Chapter 3 on cognitive validity), and second, *situational authenticity* which attempts to take into account the contextual requirements of the tasks (see Chapter 4 on context validity). Cambridge ESOL adopts an approach which recognises the importance of both situational and interactional authenticity (see Bachman and Palmer 1996 for discussion of these concepts).

The concern with situational authenticity requires writers to respond to contexts which simulate 'real life' in terms of criterial parameters without necessarily replicating it exactly. As far as possible, attempts are made to use situations and tasks which are likely to be familiar and relevant to the intended test taker. In providing contexts, the purpose for carrying out a particular Writing task is made clear, as well as the intended audience, and the criterion for success in completing the task.

Saville (2003:67) positions Cambridge ESOL examinations as follows:

> The authenticity of the tasks and materials in the Cambridge EFL examinations is often referred to as a major strength of the approach . . . The examination content must be designed to provide sufficient evidence of the underlying abilities (i.e. construct) through the way the test taker responds to this input. The authenticity of test content and the authenticity of the candidate's interaction with that content are important considerations for the examination developer in achieving high validity.

There is a strong argument for making tests as direct as possible. The more features of real-life use of language, in this case of writing, that can be built into test tasks the greater the potential for positive washback on the learning that precedes the test-taking experience and the easier it will be from the test to make statements about what students can or cannot do as regards writing. If we want an estimate of a candidate's writing ability, it seems a waste of time to be training students in ways of improving their scores on indirect tests of writing, such as multiple-choice tests of written expression as has happened in the past in some tests of writing. If the purpose is to measure writing ability, examination boards should be employing Writing tasks that encourage teachers to equip candidates with the writing abilities they will need for performing in a real-world context.