1 Introduction

This volume stands as the second in a series of volumes designed to explore the constructs underpinning the testing of English language skills. The specific focus here is on the testing of second language reading ability and the title is a companion to the first construct volume in the series, *Examining Writing*, by Shaw and Weir (2007). To some degree, *Examining Reading* covers ground already mapped out in the earlier publication in relation to testing second language writing. Where the concepts are identical there is inevitable overlap with that volume. For the most part, however, this volume reflects a novel updating of the theoretical framework for the validation of language examinations in relation to the testing of reading first outlined in Weir (2005). In addition, the volume examines the operationalisation of that framework by way of a critical evaluation of Cambridge ESOL examinations. This evaluation provides the context for the framework's exegesis as in each chapter Cambridge practice is reviewed in terms of the particular component of the framework under review.

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 to provide a coherent account of the theoretical construct on which reading examinations should be based and of the rigorous procedures that need to be followed to provide evidence concerning the various components of a test's validity. As such it is hoped that it will offer other institutions a useful framework for reviewing their own examinations/tests.

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

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.

1

Examining Reading

Firstly there is the voice of the wider academic community in Applied Linguistics and Language Testing which provides the theoretical foundation for the framework we have developed and the guiding principles on which we feel good practice should be based. In discussing each section of the framework an account is first given of contemporary thinking on the area under discussion. After we have addressed the current thinking on a particular element of the framework we examine it in detail in terms of Cambridge ESOL practice through 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, centre administrators, etc.). Sometimes the voice takes the form of case studies to exemplify particular issues; at other times it exists in quotations from or in 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 important material together from a variety of sources within the organisation relating to the operationalisation of Cambridge ESOL's examinations 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, for example see Appendices C, D and E. It reflects Cambridge ESOL's ongoing commitment to increasing transparency and accountability, and to sharing the organisation's knowledge, skills and experience with the wider language testing community.

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, environment.

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 examinations they offer are well grounded in the language ability constructs they are attempting to measure. An explicit test validation framework is required which enables examination providers to furnish comprehensive evidence in support of any claims about the soundness of the theoretical basis of their tests.

2

Examination boards and other institutions offering high-stakes tests need to demonstrate 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 develops a theoretical framework for validating tests of second language reading ability which then informs an attempt to articulate the Cambridge ESOL approach to assessment at different proficiency levels in the skill area of reading. The perceived benefits of a clearly articulated theoretical and practical position for assessing reading skills in the context of Cambridge ESOL tests are essentially twofold:

- *Within Cambridge ESOL* – this articulated position will deepen understanding of the current theoretical basis upon which Cambridge ESOL assesses different levels of language proficiency across its range of 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 versions 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 existing 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 early work of Bachman (1990) and Bachman and Palmer (1996) underpinned the adoption of the VRIP approach, as set out in Weir and Milanovic (2003), and found in various Cambridge ESOL internal documents on validity (e.g. Milanovic and Saville 1996).

We explore below how the socio-cognitive validity framework described in Weir's *Language Testing and Validation: an evidence-based approach* (2005) might contribute to an enhanced validation framework for use with Cambridge 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) might interact with each other. Reading, the construct of interest in this volume, is viewed as not just the underlying latent trait of reading ability but as the result of the constructed triangle of trait, context and score (including its interpretation). The approach adopted in this volume is therefore effectively an *interactionalist* position, which sees the reading construct as residing in the interactions between the underlying cognitive ability, the context of use and the process of scoring.

3

Examining Reading

In addition, the approach 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 parameters for distinguishing between adjacent proficiency levels are identified. The approach, building on Weir (2005), is represented pictorially 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' (Weir 2005:43). Conceptualising validity in terms of temporal sequencing is of value as it offers test developers a plan of what should be happening in relation to validation and when it should be happening.

The model 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.

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

- How are the physical/physiological, psychological and experiential characteristics of candidates catered for by this test? (focus on the *Test taker* in Chapter 2)
- Are the cognitive processes required to complete the test tasks appropriate? (focus on *Cognitive validity* in Chapter 3)
- Are the characteristics of the test tasks and their administration appropriate and fair to the candidates who are taking them? (focus on *Context validity* in Chapter 4)
- How far can we depend on the scores which result from the test? (focus on *Scoring validity* in Chapter 5)
- What effects do the test and test scores have on various stakeholders? (focus on *Consequential validity* in Chapter 6)

4

Figure 1.1 A framework for conceptualising reading test validity
(adapted from Weir 2005)
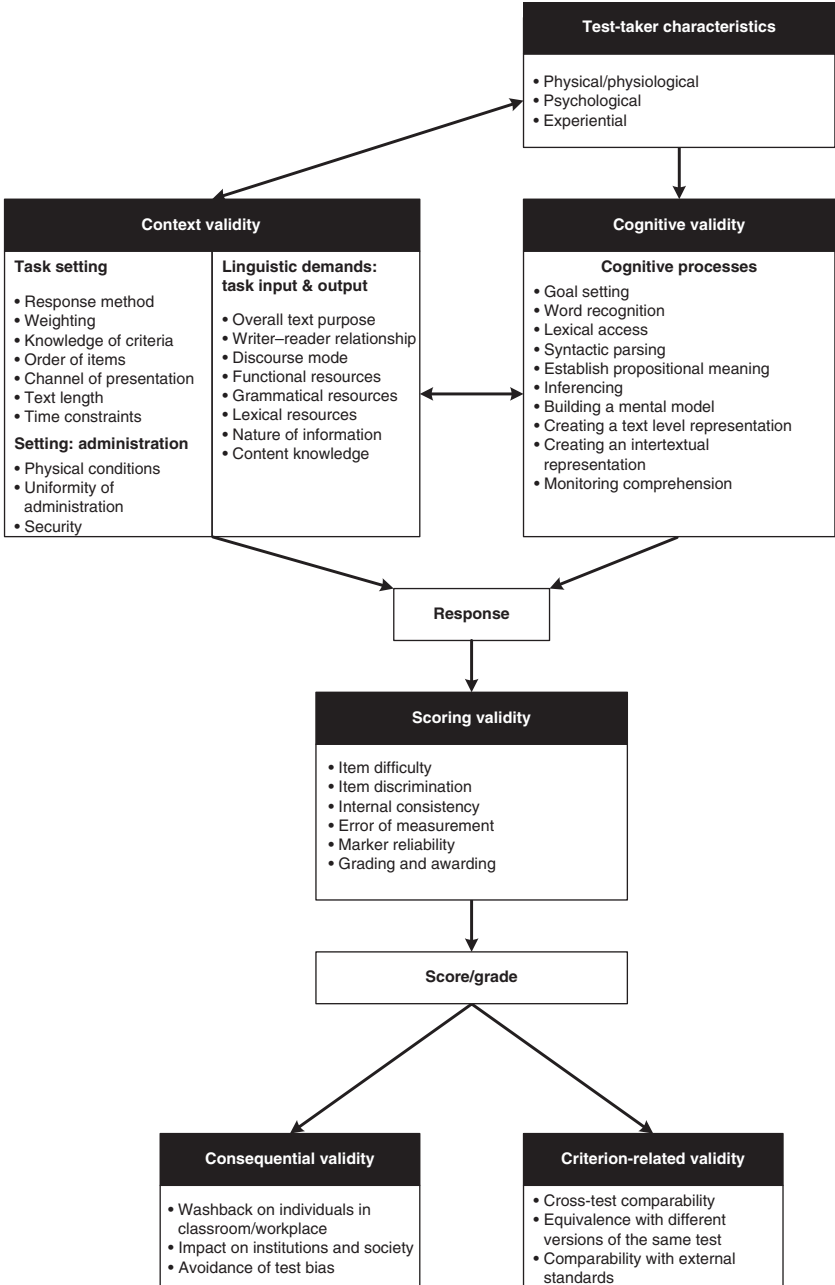
Examining Reading

- What external evidence is there that the test is fair? (focus on *Criterion-related validity* in Chapter 7)

These are the types 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 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 2005:51).

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. a test taker's educational and cultural background, experience in preparing and taking examinations as well as 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 (see Chapter 2 for detail).

Cognitive validity is established by *a priori* evidence on the cognitive processing activated by the test task before the live test event (e.g. through verbal reports from test takers), as well as through the more traditional *a posteriori* evidence on constructs measured involving statistical analysis of scores following test administration. 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 (see Chapter 3 for detail).

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 (see Chapter 4 for detail). Context validity for a reading task thus addresses the particular performance conditions, the setting under which it is to be performed (such as response method, time available, text length, order

6

of items as well as the linguistic demands inherent in the successful perform-
ance of the task) together with the actual examination conditions resulting
from the *administrative setting* (Weir 2005).

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
2005: Chapter 9, and Chapter 5 below for detail). Scoring validity accounts
for the extent to which test scores are arrived at through appropriate crite-
ria in constructed response tasks and exhibit consensual agreement in their
marking, are as free as possible from measurement error, stable over time,
appropriate in terms of their content sampling and engender confidence as
reliable decision-making indicators.

Messick (1989) argued the case for also considering *Consequential Validity*
in judging the validity of scores on a test. From this point of view it is neces-
sary 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 (see Chapter 6 below for detail). 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
issue of test bias takes us back to the *test-taker characteristics* box. The evi-
dence collected 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.

Criterion-Related validity is a predominantly quantitative and *a poste-
riori* 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 and Chapter 7 below for detail).
Evidence of criterion-related validity can come in three forms:

- Firstly if a relationship can be demonstrated between test scores and
  an external criterion which is believed to be a measure of the same
  ability. This type of criterion-related validity is subdivided 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 reading 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 another
  measure of the ability of interest for the same candidates taken some
  time after the test has been given (Alderson et al 1995).
- Demonstration of the qualitative and quantitative equivalence of
  different versions of the same test is a second source of evidence.

Examining Reading

- A third source of evidence results from linking a test to an external standard such as the Common European Framework of Reference (CEFR) through the comprehensive and rigorous procedures of familiarisation, specification, standardisation and empirical validation (Council of Europe 2003).

Validity as a unitary concept

Although for descriptive purposes the various elements of the model in Figure 1.1 are presented as being separate from each other, undoubtedly a close relationship exists between these elements, for example between context validity and cognitive validity, which together with scoring validity constitute for us 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. The interactions between, and especially within, these 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.

Focus of the volume

As a general principle, language tests should, as far as is practicable, place the same requirements on test-takers as are involved in communicative settings in non-test 'real-life' situations. This approach requires attention to both cognitive and social dimensions of communication.

A major focus of this volume is Cambridge ESOL's concern with *authenticity* which has been a dominant theme 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 Alderson 2000, Hawkey 2005, Morrow 1979, Weir 1983, 1990, 1993 and 2005). The 'real-life' 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's 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 Cognitive validity below), and second, *situational authenticity,* which attempts to take into account the contextual requirements of the

tasks (see Chapter 4 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).

Though full authenticity may be unattainable in the testing situation, as far as is possible, attempts should be made to use situations and tasks which are likely to be familiar and relevant to the intended test-taker. The concern with situational authenticity requires readers to respond to contexts which simulate 'real life' in terms of criterial contextual parameters without necessarily replicating it exactly. In this paradigm tests should be as direct as possible and, by employing tasks which activate the types of processing that characterise reading in the real life target situation, interactional authenticity is enhanced. The more features of real-life use of language, in this case of reading, 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 to extrapolate from the test to make statements about what students can or cannot do in real-life reading situations. If the purpose is to measure reading ability, examination boards should be employing reading tasks that encourage teachers to equip candidates with the reading abilities they will need for performing in a real-world context.

Cambridge ESOL's Main Suite examinations offer a picture of how reading ability is measured by the examination board across a broad language proficiency continuum. Its five levels correspond to equivalent levels of the Association of Language Testers in Europe (ALTE) and the Common European Framework of Reference (CEFR). These levels are compatible and correspond to the levels of language ability familiar to English language teachers around the world, i.e. from beginner to advanced. The relationship between Cambridge ESOL levels, ALTE levels and the CEFR levels is discussed in detail in Chapter 7. However, for initial orientation the reader is referred to Table 1.1 below for an overview of ALTE 'Can Do' statements and Table 1.2 which provides CEFR A2 to C2 illustrative scales for reading (see Jones 2002 and Chapter 7 for details of a Cambridge ESOL project which linked the ALTE levels to the CEFR). We then provide a description of Main Suite levels in terms of what materials successful candidates can handle and what they are expected to be able to do at each of the five levels (see Table 1.3).

When considering the ALTE Table (1.1) and the CEFR Table (1.2), the reader may feel that the distinctions between adjacent levels are not always clear and the characterisations on occasion imprecise. It is the intention of this volume to try and improve on these descriptions by clarifying the underlying theoretical construct of Reading at CEFR levels A2 to C2 and by a close examination of Cambridge ESOL practice to specify more precisely, where possible, any differences between adjacent levels in terms of a range of contextual and cognitive parameters.

Although the Main Suite, a set of General English examinations, forms

Examining Reading

Table 1.1 ALTE 'Can Do' statements

ALTE Levels (CEFR Levels)	ALTE 'Can Do' Statements			
	Overall general ability	Social & Tourist typical abilities	Work typical abilities	Study typical abilities
ALTE Level 5 (C2: Mastery)	CAN understand documents, correspondence and reports, including the finer points of complex texts	CAN (when looking for accommodation) understand a tenancy agreement in detail, for example, technical details and the main legal implications	CAN understand reports and articles likely to be encountered during his/her work, including complex ideas expressed in complex languages	CAN access all sources of information quickly and reliably
ALTE Level 4 (C1: Effective Operational Proficiency)	CAN read quickly enough to cope with an academic course, to read the media for information or to understand non-standard correspondence	CAN understand complex opinions/ arguments as expressed in serious newspapers	CAN understand correspondence expressed in non-standard language	CAN read quickly enough to cope with the demands of an academic course
ALTE Level 3 (B2: Vantage)	CAN scan texts for relevant information, and understand details, instructions or advice	CAN understand detailed information, for example a wide range of culinary terms and abbreviations in accommodation advertisements	CAN understand most correspondence, reports and factual product literature he/she is likely to come across	CAN scan texts for relevant information and grasp main point of text
ALTE Level 2 (B1: Threshold)	CAN understand routine information and articles, and the general meaning of non-routine information within a familiar area	CAN understand factual articles in newspapers, routine letters from hotels and letters expressing personal opinions	CAN understand the general meaning of non-routine letters and theoretical articles within own work area	CAN understand basic instructions and messages, for example, computer library catalogues, with some help
ALTE Level 1 (A2: Waystage)	CAN understand straightforward information within a known area, such as on products and	CAN understand straightforward information, for example labels on food, standard menus,	CAN understand most short reports or manuals of a predictable nature within his/her own area	CAN understand the general meaning of a simplified textbook or article, reading very slowly

10