1 Introduction

Lynda Taylor
Consultant to Cambridge ESOL

The context for the volume

Examining Speaking stands as the third in a series of volumes designed to explore the constructs underpinning the testing of English language skills. The specific focus in this volume is on the testing of second language speaking ability. The title is a companion to two earlier construct-focused volumes in the series, *Examining Writing* by Shaw and Weir (2007) and *Examining Reading* by Khalifa and Weir (2009). A fourth volume – *Examining Listening* – is in preparation at the time of writing and its publication will complete a set of four volumes focusing on the four skills as they have been traditionally conceptualised and operationalised by the language testing and assessment community over many years.

This fourfold categorisation of language proficiency (i.e. according to the skills of reading, writing, listening and speaking) has been adopted as the organising principle behind the series because it continues to occupy a central role in the activities of examination boards and other language test providers. However, some consideration is given to integrated skills testing, for example, reading into writing in summary tasks in the reading volume, and reading as input into writing tasks in the writing volume. Some descriptive frameworks choose to compartmentalise overall language proficiency according to dimensions or skill-sets other than the traditional quartet, e.g. enabling skills (such as lexis and grammar, etc.), or they focus instead on integrated skills (listening/speaking, reading into writing). Interestingly, the *Common European Framework of Reference for Languages: Learning, teaching, assessment* (Council of Europe 2001) proposes several different ways of categorising language proficiency: one approach subdivides overall proficiency into Speaking/Writing/Understanding (with Understanding being used to cover Reading and Listening); an alternative approach conceptualises proficiency according to Receptive/Productive/Interactive dimensions (in this case Speaking can be either Productive or Interactive); and a third approach adds a further skill of Mediation. In light of this, the four volumes, each with their dominant focus on a single language skill, are not intended to offer the definitive or last word on approaches to describing language proficiency for assessment purposes. As the language testing profession continues to reconceptualise and expand its

understanding of the complex interaction of skills making up language proficiency, it is always possible that additional construct-oriented volumes will be added to the *Studies in Language Testing* series in the future.

Given its function as part of a coherent set of skills-focused volumes, *Examining Speaking* not surprisingly covers some of the ground already mapped out by its predecessors, *Examining Writing* and *Examining Reading*. A strong family resemblance will be discernible between the volumes, in terms of both format and approach. The theoretical framework for validating language examinations first outlined in Weir (2005a) remains the springboard for reflecting upon our understanding and conceptualisation of the speaking ability construct for assessment purposes. The Cambridge ESOL examinations are once again taken as the practical context for undertaking a critical evaluation of Speaking tests ranging across different proficiency levels, enabling us to examine how the theoretical framework for validation can be operationalised in practice and with what outcomes. As in previous volumes, each chapter closely scrutinises Cambridge practice in terms of the particular component of the framework under review. Although each volume replicates the approach of its predecessors to some degree, it also seeks to build upon and extend the earlier work reported in the series, bringing fresh and novel insights into the process of construct definition and operationalisation for the particular skill of interest. It thus allows Weir's original (2005a) theoretical framework to be developed and refined in light of the experience of applying it in practice.

It is only appropriate at this point to acknowledge the existence of other important frameworks and models that are available to language testers and examination boards. These include Evidence-Centered Design (ECD) as proposed by Mislevy, Steinberg and Almond (2002, 2003; see also Mislevy, Almond and Lukas 2003), and Assessment Use Argument (AUA), as set out by Bachman (2005) and Bachman and Palmer (2010). Other test providers have found these to be an accessible and fruitful way of guiding their practical test design and validation, as demonstrated by Chapelle, Enwright and Jamieson (2008). However, Cambridge ESOL has found the socio-cognitive model, first offered by Weir in 2005 and later refined through the experience of applying it to operational tests, to match well with the kinds of tests the examination board produces, addressing the validation questions that arise and providing some of the answers that are needed. The model has proved to be both theoretically sound and practically useful over a number of years in relation to a variety of different examinations produced by Cambridge ESOL and for this reason is used as the framework for description and analysis in this and previous volumes.

The intended audience for the volume

The intended audience for the volume is primarily the constituency of professional language testers who are directly involved in the practical assessment

of second language speaking ability. This growing constituency around the world typically includes staff working for examination boards and testing agencies at regional, national and/or international level, as well as those working within education ministries whose remit is to advise on the development of assessment policy or the implementation of examination reform programmes. In an age where public institutions such as examination boards and related organisations are increasingly called to account for how and why they design and administer their tests in the way they do, issues of openness and transparency are of growing importance. We believe that the approach outlined and exemplified in this volume offers one way of placing in the public domain the rationale and evidence in support of testing policy and practice. By sharing one examination board's expertise and experience in this way, we hope that other institutions and test providers will be encouraged and enabled to review and reflect upon their own testing theory and practice, and thus engage in a similar exercise in public accountability for their own assessment products.

The volume is also directly relevant to the academic language testing and assessment community, i.e. researchers, lecturers and graduate students, especially those with a specific interest in assessing spoken language ability. While the overview of the theoretical and empirical research should be of obvious and immediate value to them, members of the academic language testing community will hopefully find the detailed description and discussion of operational language testing practices just as useful. The complex practical constraints facing an examination board can sometimes mean that experimental research findings are not immediately or readily applicable to large-scale testing activities. Operational testing, as opposed to language testing research, often has to concern itself with far more than just the issues of construct definition and operationalisation, assessment criteria and rating scale development. Large-scale commercial tests conducted on an *industrial* scale, such as those offered by Cambridge ESOL and similar agencies, are usually located within a complex ecology comprising multiple, interacting factors, many of which are simply not present or relevant in more academically oriented language testing research endeavours. Such factors include sustainability issues to do with test production, delivery and processing systems; practical issues concerning test timing, security, cost, accessibility; organisational issues relating to personnel (e.g. developing and sustaining the rater cadre) or management (e.g. the revision of an existing test, or development of its replacement). This is particularly true for both the direct and the semi-direct testing of L2 speaking ability, for which practicality and sustainability are core considerations, as will become apparent later in this volume. Hopefully, the explication of theory and practice presented in this volume will contribute to a broader and deeper understanding of the issues for all of us who are involved in the assessment of speaking.

Examining Speaking

There will undoubtedly be other readers for whom certain sections, if not the whole, of this volume will be of interest and relevance, perhaps because they are involved in teaching and assessing L2 spoken ability, or because they are preparing learners to take one or more of the Cambridge ESOL examinations. Such readers include English language teachers, curriculum developers, course book writers and other materials developers. Although, for obvious reasons, the primary focus of this volume is on the testing of English as a second language, we hope that some of the theory, principles and practice that are explored and explained in the volume will prove helpful to teachers and learners of languages other than English as a second/foreign language, and especially to teachers of those less commonly taught languages for which assessment theory and testing practice are still in the early stages of development and remain relatively under-resourced.

In addition to the audiences highlighted above, we anticipate that this volume will be of direct interest to the vast community of English language professionals who, in one way or another, are personally involved with the Cambridge ESOL Speaking tests. They include the hundreds of test materials writers who draft and edit Speaking test tasks and rubrics for the multiple proficiency levels, as well as the many thousands of Speaking test personnel (speaking examiners (SEs)) around the world, working in teams alongside Team Leaders (TLs), Regional Team Leaders (RTLs) and Professional Support Leaders (PSLs) to deliver the Speaking tests and to assess test takers' performances as fairly as possible. The Cambridge ESOL examinations could not function as successfully as they do without the expertise and dedication of this professional cadre worldwide. Indeed, these specialists have made their own contribution to the development of this volume as this chapter will make clear.

Finally, in a globalised world where the testing and assessment of second, third or additional language skills are steadily moving centre stage in education and society, this volume is offered as a contribution towards the promotion of assessment literacy. Language tests and the scores they generate are increasingly used across contemporary society worldwide: within education, from primary age to higher education; in employment contexts, from the professional registration of health professionals, to health and safety issues in the catering or construction industry; and, more controversially, in migration and citizenship policy and practice around the world. These trends mean that there are not simply more people taking tests. Growing numbers of people are directly involved in selecting or developing tests and in using test scores for decision-making purposes. In practice, they often find themselves doing this without much background or training in assessment to equip them for the role. They include classroom teachers tasked with designing or delivering standardised tests to evaluate their pupils' progress, tests that are sometimes then used to hold teachers and schools accountable for that progress, or the apparent lack of it. A similar burden of expectation is laid on staff in

university admissions, in professional bodies and in immigration agencies to know what tests measure and what test scores mean, and to understand how to integrate them into their complex, usually high-stakes decision-making processes. The language testing community is in a position, indeed it has a moral obligation, to encourage the sharing of the core knowledge, skills and understanding that underpin good quality assessment as widely and accessibly as possible for the benefit of all. (For a fuller discussion of the importance of assessment literacy and approaches to its development, see Taylor 2009a.)

In summary, then, this volume is offered as a rich source of information for a wide variety of audiences on multiple aspects of examining L2 speaking ability.

The purpose of the volume

As explained above, *Examining Speaking* is one of a series of *construct*-oriented volumes focusing upon the four language skills of writing, reading, speaking and listening. The genesis of the series lies partly in a close collaboration that developed in the early 2000s between, on the one hand, applied linguistics and language testing specialists at Cambridge ESOL, and, on the other, Professors Cyril J Weir and Barry O'Sullivan, both of whom were at that time working at the University of Roehampton, Surrey. There was on their part a shared interest and enthusiasm with Cambridge ESOL in finding ways to explore more systematically the nature of construct validity in language testing and assessment, and especially to bridge the gap between research and practice, between theoretical construct definition and applied construct operationalisation in relation to real-world language testing, particularly the sort of large-scale language testing undertaken by examination boards and agencies such as Cambridge ESOL. Initial discussions led to the conceptualisation of a long-term project to research and draft a series of documents (e.g. position papers, research reports, and published monographs) which would describe and reflect upon the theory and practice of assessment and how this is operationalised in the Cambridge ESOL examinations, with particular reference to the multiple proficiency levels of the General English suite of tests, traditionally referred to as the Cambridge Main Suite (MS). It was envisaged this project might include the publication of a set of academic volumes within the *Studies in Language Testing* series.

A major motivation for embarking on such a project was the growing expectation in the public domain, both nationally and globally, for examination boards and other test providers to be transparent and accountable in what they do, especially in terms of the standards to which they adhere, the quality and validity claims they make for their products, and the provision of theoretical and empirical evidence in support of these claims. This external, public expectation emerging within wider society was paralleled by

Examining Speaking

a growing sense within the language testing and assessment profession itself of the need to develop its own *professional ethic*. It is in the light of this awareness that the field of language testing and assessment has undergone a process of increasing professionalisation over recent years. An abundance of quality standards, ethical codes and guidelines for good testing practice has been embraced or generated by language testers, in many cases touching upon matters that extend well beyond a test's purely technical qualities. Examples include the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices 2004) and the AERA/APA/NCME *Standards* (1999), both of which concern assessment in general. There also exist language testing specific codes such as the ALTE *Code of Practice* (1994), the ILTA *Code of Ethics* (2000), the EALTA *Guidelines for Good Practice* (2006) and the ILTA *Guidelines for Practice* (2007). Professional associations of language testers and testing organisations were established during the 1990s at national, regional and international level, including the Japan Language Testing Association (JLTA), the Association of Language Testers in Europe (ALTE), the International Language Testing Association (ILTA), and the European Association of Language Testing and Assessment (EALTA). Kunnan (2004, 2008) reflects on the ethical milieu that emerged for language testers, prompted by various factors in the professional field such as: demands for accountability and responsiveness to clients; increased use of language tests and types of delivery methods; use of new measurement and other analytical techniques; and expanded concepts of validity (i.e. Messick 1989, 1996). Kunnan suggests the language testing community responded to these factors by developing explicit standards and codes for its community and practice, and that its professional ethic continues to evolve informed by the wider literature in ethics and moral philosophy.

An ethic that embraces openness, transparency and accountability is essential given that examination boards and other testing institutions offer assessment tools whose use has both direct and indirect consequences for education and wider society. Such consequences may be high stakes, influencing an individual's life chances, the formulation of public policy or the shaping of attitudes in society. Messick's re-conceptualisation of validity, which brought together traditional validity concerns but also added value implications and social consequences as essential facets, undoubtedly contributed to growing awareness of the consequences of testing, intended and unintended, positive and negative; and this trend is clearly illustrated in the wealth of research literature published over recent years on the theory and practice of language testing washback and impact (see, among others, Alderson and Wall 1996, Cheng 2005, Cheng, Watanabe and Curtis 2004, Green 2007, Hawkey 2006, Kunnan 2000, Wall 2005, Wall and Horak 2006, 2008), as well as other publications discussing the role of testing in education and society (see, for example, McNamara and Roever 2006, Shohamy 2001, 2008, Spolsky 2008).

Transparency and accountability are particularly important for testing organisations that offer multiple tests targeted at different proficiency levels, domains of language use or groups of language users. In this context, examination providers need recourse to an explicit and appropriate methodology for describing, analysing and comparing their tests in a systematic and comprehensive manner so that test users can clearly understand the features of each testing option available to them and can decide appropriately on their selection and use.

Cambridge ESOL has never subscribed to a philosophy of 'one size fits all' where English language assessment is concerned. Over time, in response to changing trends in language learning and to evolving market demands and opportunities, the examination board has developed a wide range of assessment products that include: tests at *different proficiency levels* (e.g. KET, PET, FCE, CAE, CPE); tests involving a *multi-skills package* (e.g. IELTS) and tests which are *modular* (e.g. ESOL Skills for Life); tests across *different language domains* (e.g. General English, Academic English, Business English); tests for *teachers* of English (e.g. CELTA, DELTA) and tests for *young learners* of English (e.g. YLE); tests in *pencil and paper mode* (e.g. Standard BULATS) and tests in *computer mode* (e.g. BULATS Online Courses); tests for *certificated* use and tests for *institutional* use.

The development and promotion of a variety of testing instruments places an obligation upon the test producer to be able to clearly demonstrate how they are seeking to meet the demands of validity in each product and, more specifically, how they actually operationalise criterial distinctions, not only between tests offered at different *levels*, i.e. on the *vertical* proficiency continuum, but also between alternative testing *domains*, *formats* and *modes*, i.e. along a *horizontal* axis.

To be able to do this requires some sort of methodology for analysing and describing the component validity features of any test, as well as for constructing an interpretative framework of reference within which multiple tests and their respective validity features can be explicitly presented and co-located. The use of such a methodology has the potential to achieve two significant and beneficial outcomes for the language testing world. First, it should enable test *producers* to assemble and present, with some degree of transparency and coherence, the validation evidence and arguments in support of quality claims made for each of their tests so that these can be scrutinised and evaluated. Secondly, it should serve as a means of communication, assisting test *users* to understand better the nature of the testing tools available to them and aiding them in decisions about which test or tests best suit a given purpose and context of use. Transparency and coherence are stated aims underpinning the development of the Common European Framework of Reference (CEFR), first released by the Council of Europe in draft form for consultation in 1996, and formally published in 2001, as

7

explained by one of its original authors Brian North (2008:21). North also expressed the hope that the CEFR would 'establish a metalanguage common across educational sectors, national and linguistic boundaries that could be used to talk about objectives and language levels', as well as 'providing encouragement to practitioners in the language field to reflect on their current practice' (2008:22). Cambridge ESOL's construct volumes embody similar aspirations and can thus be seen within a broader historical frame, especially, though not exclusively, within the European language education context, in which there has been an increasing focus on the need to analyse and describe language proficiency for the purposes of learning, teaching, and assessment.

The theoretical framework for validating language examinations first outlined in Weir's *Language Testing and Validation: An evidence-based approach* (2005a) offered Cambridge ESOL a potential approach and methodology for undertaking such an enterprise in relation to its own examinations. Drawing upon theoretical and empirical research in the field, it provided a useful conceptual heuristic for identifying core features in the process of improved construct definition for the tests. More importantly, perhaps, it also offered the hope that it could be proactively applied to operational tests, i.e. to the *test-in-practice* rather than just the *test-in-theory*. It was anticipated that a socio-cognitive framework for validating examinations, as expounded by Weir (2005a), would permit a systematic and comprehensive critical evaluation of construct definition *and* operationalisation, and ideally furnish explicit evidence, both theoretical and empirical, to support claims about the usefulness of the Cambridge ESOL tests. The socio-cognitive approach to test validation resonated strongly with the thinking and practice on test development and validation which had been emerging in Cambridge ESOL during the 1990s, 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 1992a, 1996a). 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. Speaking, the construct of interest in this volume, is viewed as not just the underlying latent trait of speaking 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 speaking construct as residing in the interactions between the underlying cognitive ability, the context of use and the process of scoring, as discussed by Weir (2005a), as well as in an earlier internal paper by O'Sullivan and Weir (2002), originally commissioned by Cambridge ESOL to explore some of the research issues involved in L2 speaking assessment.

Like its predecessors *Examining Writing* and *Examining Reading*, therefore, this volume develops a theoretical framework for validating tests of second language speaking ability, a framework which then informs an attempt to articulate and evaluate the Cambridge ESOL approach to assessing L2 speaking skills. The perceived benefits of a clearly articulated theoretical and practical position for assessing speaking skills in the context of the Cambridge ESOL examinations 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 forms and tasks.

• *Beyond Cambridge ESOL* – it will communicate in the public domain the theoretical basis for the tests and hopefully provide a more clearly understood rationale for the way in which Cambridge ESOL operationalises this in its tests. In addition, it may provide a suitable framework for others interested in validating their own examinations, offering a principled basis and a practical methodology for comparing language examinations across the proficiency range. It therefore adds to the range of frameworks and models now available to test developers for analysing and describing the qualities of their tests and for guiding their research and validation activity.

The focus of the volume

The intention, then, in this volume is to apply a theoretical framework for validating tests of second language speaking ability in order to examine, articulate and evaluate the approach to assessing L2 speaking skills adopted by Cambridge ESOL. The board's suite of examinations in General English (the Main Suite) offers a useful picture of how speaking ability is measured across a broad language proficiency continuum, i.e. from beginner to advanced. Its five levels correspond to equivalent levels of ALTE and of the Common European Framework of Reference (CEFR). The levels reflect the levels of language ability familiar to English language teachers around the world, that have been described as 'natural levels' (North 2006:8), not in the sense that they are themselves naturally occurring phenomena in either language acquisition or learning, but rather in the sense that language teachers and educators, especially in ELT, gradually came to perceive them as *useful* curriculum and examination levels. 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 for an overview of

Examining Speaking

ALTE Can Do statements for Listening/Speaking and to Table 1.2 which
provides CEFR A1 to C2 performance level descriptors for Speaking. In their
development, the ALTE Can Do statements were originally organised into
three general areas: Social and Tourist; Work; and Study. Each of these areas

Table 1.1 ALTE Can Do statements for Listening/Speaking

CEFR Levels (ALTE Levels)	Listening/Speaking Can Do Statement			
	Overall General Ability	Social and Tourist typical abilities	Work typical abilities	Study typical abilities
C2: Mastery (ALTE Level 5: Good User)	CAN advise on or talk about complex or sensitive issues, understanding colloquial references and dealing confidently with hostile questions.	CAN talk about complex or sensitive issues without awkwardness.	CAN advise on/handle complex delicate or contentious issues, such as legal or financial matters, to the extent that he/she has the necessary specialist knowledge.	CAN understand jokes, colloquial asides and cultural allusions.
C1: Effective Operational Proficiency (ALTE Level 4: Competent User)	CAN contribute effectively to meetings and seminars within own area of work or keep up a casual conversation with a good degree of fluency, coping with abstract expressions.	CAN keep up conversations of a casual nature for an extended period of time and discuss abstract/cultural topics with a good degree of fluency and range of expression.	CAN contribute effectively to meetings and seminars within own area of work and argue for or against a case.	CAN follow abstract argumentation, for example the balancing of alternatives and the drawing of a conclusion.
B2: Vantage (ALTE Level 3: Independent User)	CAN follow or give a talk on a familiar topic or keep up a conversation on a fairly wide range of topics.	CAN keep up a conversation on a fairly wide range of topics, such as personal and professional experiences, events currently in the news.	CAN take and pass on most messages that are likely to require attention during a normal working day.	CAN give a clear presentation on a familiar topic, and answer predictable or factual questions.

10