1 An overview of the influences on English language testing in the United Kingdom 1913–2012

Cyril J Weir

University of Bedfordshire

> . . . who will write the more general, more fluid, but also more determinant history of the 'examination' – its rituals, its methods, its characters and their roles, its play of questions and answers, its systems of marking and classification? For in this slender technique are to be found a whole domain of knowledge, a whole type of power.
> *Discipline and Punish: The Birth of the Prison* (Foucault 1979: 185)

Introduction

In recent years assessment issues have assumed an increased importance in the economic, educational and socio-political affairs of society. Spolsky (2008a: 297) argues that by the 21st century 'testing has become big business', and Shohamy (2008:xiv) points to 'the societal role that language tests perform, the power that they hold, and their central functions in education, politics and society'. A significant role for testing language proficiency can be seen *inter alia* in migration and citizenship policy and practice, the professional registration of those involved in the provision of health care, appointment and promotion in business, industry and commerce, the certification of air traffic and maritime personnel, and entry to tertiary level education. Such uses testify to the critical gatekeeping function that language assessment fulfils in contemporary society.

As the power of tests and the potential for their misuse/abuse grows, assessment literacy in our society seems more important than ever. Taylor (2009:29) argues that narrative accounts which chronicle testing developments over time may have an important role to play in fostering this:

> They contextualize the practice of language testing as a socially constructed and interpreted phenomenon, rather than treating it primarily as a pseudoscientific endeavour that is removed and isolated from human individuals and social values. It may well be that popular adaptations of this narrative, storytelling approach will prove a more effective means of developing assessment literacy among the wider stakeholder community in the future.

1

Measured Constructs

Despite the importance of language assessment in the United Kingdom, we still lack a satisfactory account of its historical development in this country. In *A History of English Language Teaching*, Howatt and Widdowson (2004:332) acknowledged that they had not given English language testing 'the prominence it deserves'. Spolsky's (1995a) authoritative work on language testing, *Measured Words*, for the most part covered the development of English language testing in the United States with only partial reference to events on this side of the Atlantic. In this volume we seek to make good this discrepancy by adopting a mainly British perspective which focuses on Cambridge English language examinations over the course of the last century. This takes us from a small cottage industry in 1913, with the Certificate of Proficiency in English (CPE) administered to three students in one London centre, to the big business of a leading international examining board in 2012. Cambridge now offers multiple English language examinations at different levels, across different domains, to over 4 million candidates per annum, in 2,700 authorised centres across 130 countries.

In delimiting our study to Cambridge English language examinations we are mindful of the advice of Stern (1983:83) that: 'by selecting a restricted field historians have a better chance of discovering and analyzing a manageable body of data'. Accordingly the centre of attention in our survey 1913–2012 is on English language testing developments in one examining board in the United Kingdom. This will enable us in Stern's terms to examine English language testing more rigorously, both synchronically at a given stage in history in a social and educational context, and diachronically over a 100-year period.

We will for the most part further limit our study to those general English examinations for non-native speakers developed in Cambridge from 1913–2012. We will concentrate on the Certificate of Proficiency in English (CPE) 1913–2012 and the Lower Certificate in English (LCE) 1939–1975, rebranded as the First Certificate in English (FCE) 1975–2012. The history of these particular examinations offers us an accessible and manageable perspective and critically it enables us to make comparisons in relation to the same examinations over an extended period of time. Such a focus will enable us to trace *continuity* and *innovation* in the measurement of language constructs in one examination board over an entire century.

Initially these were known as English as a Foreign Language (EFL) examinations but more recently as examinations in English for Speakers of Other Languages (ESOL). For ease of reference, especially when talking about them collectively, we will use the generic term Cambridge English examinations throughout. The body responsible for running the English language examinations was initially the University of Cambridge Local Examinations Syndicate (UCLES), then UCLES EFL and more recently Cambridge ESOL. We will maintain these distinctions when referring specifically to the examination board itself.

At various points in our history we will compare what was happening in the United Kingdom with developments in the United States, the two world leaders in the field of language testing in the last century as in this. Although English Language testing is now similarly informed by most aspects of construct validity in both countries, this was not always the case and the path testing was to take differed markedly in each. The reasons behind these contrasting journeys can be found in the prevailing, socio-economic contexts in the two countries, but the differing approach in the United Kingdom also reflects European legacies from the past in both theoretical and practical approaches to language teaching and assessment. A comparison of the two traditions in the United States and United Kingdom is informative and helps us understand some of the differing, if less clear-cut, emphases that are still present in approaches to theory and practice in language assessment in these two countries.

Our historical approach signifies the importance of establishing a 'big picture' to serve as a holistic frame of reference into which present phenomena can be fitted. For the language testing community to fully understand its present practice, we feel it needs to appreciate how its past has been shaped.

We first establish the main themes of our study viz *language constructs*, and the impact of *external forces* and *language teaching practice* on how these constructs were measured in Cambridge English examinations over the period 1913–2012.

Language constructs

Alan Davies (1984:68) wrote in the first issue of the journal *Language Testing*:

> . . .in the end no empirical study can improve a test's validity. . . What is most important is the preliminary thinking and the preliminary analysis as to the nature of the language learning we aim to capture.

The construct(s) Cambridge sought to measure in English language test tasks are the focal point of our story. From a 21st century testing perspective, we consider a construct as not just the underlying, latent trait of a particular ability measured by a test task but a combination of the trait, the context in which the task is performed and the criteria used for scoring performance on that task, together with the interpretation of the resulting scores. This approach is effectively an *interactionalist* position, which sees the construct as residing in the interactions between an underlying cognitive ability, a context of use and a process of scoring. Cambridge ESOL's current approach to construct validation might thus be labelled *socio-cognitive* in that the abilities to be tested are demonstrated by the cognitive processing of the candidate; equally, the use of language in performing tasks is viewed as a social rather than a purely linguistic phenomenon.

Measured Constructs

The socio-cognitive framework utilised to describe the measurement of constructs in this volume builds on Weir (2005a) and represents a unified approach to establishing the overall validity of a test (see also Field 2013). It addresses the three central aspects of construct validity:

1. *Cognitive validity*. Do the cognitive processes required to complete test tasks sufficiently resemble the cognitive processes a candidate would normally employ in non-test conditions, i.e. are they *construct relevant* (Messick 1989)? Are the range of processes elicited by test items sufficiently comprehensive to be considered representative of real-world behaviour i.e., not just a small subset of those which might then give rise to fears about construct *under-representation*? Are the processes appropriately calibrated to the level of proficiency of the learner being evaluated?

2. *Context validity*. Are the characteristics of the test task an adequate and comprehensive representation of those that would be normally encountered in the real life context? Are they appropriately calibrated to the level of proficiency of the learner being evaluated?

3. *Scoring validity*. How appropriate and comprehensive are the criteria employed in evaluating test output? How well calibrated are they to the level of proficiency of the learner being evaluated? Also included here is the traditional concern with *reliability*: how far can we depend on the scores which result from applying these criteria to test output?

Scoring validity is used as the superordinate term here in preference to the traditional term reliability as it embraces more than just the statistical consistency of the scores themselves. In particular, the *criteria* used in the assessment of test output are seen as a critical and integral part of the construct being measured. The term scoring validity emphasises its part in a unified conception of validity as compared to the narrower term reliability, the frequent use of which in conjunction with the term validity (as in 'validity and reliability') risks conveying the idea of separateness.

Language test constructs were never as explicitly or comprehensively specified in the early tests of the 'pre-scientific period' (Spolsky 1978), the 'Garden of Eden' as it was termed by Morrow (1979), when the University of Cambridge Local Examinations Syndicate (UCLES) examinations started back in 1913. As Taylor (2011a: 21) points out: 'We cannot be definitive about what was in the minds of the original CPE test developers and it is probably unlikely that the test developers of a century ago worked with the terms and definitions that are familiar to language testers today.'

While language examinations under UCLES, UCLES EFL and later Cambridge ESOL must have been intended to measure a trait or 'construct' (Saville 2003), for much of the period under review we have no direct evidence as to what test developers thought the underlying language construct(s) they

4

intended to measure were. In the first half of our history, there is little docu-
mented evidence available on the underlying basis for test construction, no
suggestion of systematic post hoc analysis of test outcomes, nor any evident
concern with the use made of test results in society. The language constructs
underlying examinations in these early days have to be interpreted largely
from the test papers themselves and what little supporting documentation is
still available. Examination revisions in recent times are better documented,
as Hawkey (2005, 2009), and Weir and Milanovic (2003) bear testimony
to. The early period thus presents a greater challenge for the historian as it
requires more interpretation as to what the test developers intended to test
than is necessary for recent practice.

An explicit definition of test construct in Cambridge ESOL is in fact a
relatively recent phenomenon. The influence of the *Cambridge-TOEFL
Comparability Study* in the late 1980s (see Bachman, Davidson, Ryan and
Choi 1995) was a catalyst for this, with its promotion of a broad construct
of communicative language ability, first proposed by Canale and Swain
(1980) in their seminal paper *Theoretical bases of communicative approaches
to second language teaching and testing* and later more fully specified by
Bachman (1990) and Bachman and Palmer (1996).

Socio-economic factors such as the move towards European standardisa-
tion, occasioned by the growth of the European Community from the 1970s
onwards, required a more explicit, transparent specification of linguistic
performance at differing levels of ability from beginner to advanced, and
together with the Comparability Study served to encourage greater attention
to test constructs and construct validation at Cambridge ESOL.

A noticeable exception to this general lack of attention to construct was a
concern with rater reliability (an important element of scoring validity) which
taxed examination providers much earlier in our history than either context
or cognitive validity. The issue of reliability was aired at an early stage by
Edgeworth (1888), Burt (1921) and Hartog, Rhodes, and Burt (1936) in the
United Kingdom, and by Thorndike (1904, 1911 and 1912) and Wood (1927)
amongst others in the United States. The reliability aspect of scoring validity
was to be the dominant validity paradigm for testing in the United States,
but it was not to be of particular concern to those testing English language at
Cambridge for most of the 20th century, until the establishment of a profes-
sional English language testing cadre there from 1988 onwards.

With an increased public expectation of transparent and explicit test spec-
ification in the late 20th century, a broader conceptualisation of construct
validity (i.e. qualitative as well as quantitative) was seen as necessary. The
current demand from stakeholders in high-stakes examinations, for example
learners, employers, receiving institutions, professional bodies (see Khalifa
and Weir 2009:177 for a full listing of all potential test stakeholders), is for
transparent, comprehensive, evidence-based answers to questions on the

Measured Constructs

language construct(s) being measured in the tests they use. Test providers need to satisfy stakeholders' expectations concerning the comparability of the constructs measured by each test version in terms of both cognitive and contextual, as well as scoring validity (Weir and Wu 2006).

The publication of four 'constructs' volumes in the *Studies in Language Testing* (SiLT) series in the early 21st century, almost 100 years after the first CPE examination, was one response to this modern day imperative for comprehensive specification and transparency (see Shaw and Weir 2007, Khalifa and Weir 2009, Taylor (Ed.) 2011 and Geranpayeh and Taylor (Eds) 2013). These volumes, building on the earlier work of Weir (2005a), further developed and elaborated socio-cognitive frameworks for specifying the various elements of construct validity in each of the broad skills areas of writing, reading, speaking and listening and then applied the parameters within these frameworks to an analysis of current Cambridge English examinations.

In Chapters 2–5 of this volume we first provide a general overview of how new ideas in teaching, applied linguistics and testing, as well as external societal forces, acted as drivers for change in the way each construct (reading, writing, speaking and listening) was conceived and measured at various points in the last century. We then look more closely at any surviving evidence from the Cambridge archives relating to the measurement of these constructs in its English language examinations and describe in greater detail how approaches changed along the way from 1913–2012. To analyse accurately and precisely the nature of these specific changes in Cambridge's approach we needed a conceptual framework to help us understand how each construct was viewed at a given point in time.

The socio-cognitive frameworks in the four 'constructs' volumes offer a useful heuristic for analysing the key features of the constructs underlying test tasks (Weir and O'Sullivan 2011). They enable us to draw on workable categories of description for informing our understanding of the changing nature of the constructs being measured by Cambridge tests during the 100 years under review in this volume. These 21st century frameworks will not be used to critique and criticise tests developed in earlier times, rather, their function is to provide a useful lens through which features of specific tests – present or past – can be discussed. Additionally they will serve as consistent frames of reference for tracking and discussing the changes to constructs in these tests over time. We are not intent on privileging one test format over another but on establishing the ways in which the measurement of a particular construct changed as test formats came into and went out of fashion.

As well as analysing the nature of the constructs being measured at various stages, we also need to consider the effects upon test design of the wider social context and of shifts in pedagogical practice, if we are to understand fully why any specific changes came about.

The impact of external forces on language testing

Spolsky (1990a:159) rightly reminds us of the important connection of developments in testing with:

> . . .external, non theoretical, institutional social forces, that on deeper analysis, often turn out to be a much more powerful explanation of actual practice. . . A clearer view of the history of the field will emerge once we are willing to look carefully at not just the ideas that underlie it, but also the institutional, social and economic situations in which they are realized.

Individuals and ideas certainly have an impact on test development as we will see but so too do institutions, government national and regional, and social and economic forces. As far as possible in this volume, we will try to locate English Language Teaching and associated testing developments in their wider social and economic contexts. By doing so, we hope to identify the influence of both the prevailing socio-economic conditions as well as the human ideas, experiences and practices that influenced both testing at Cambridge and teaching in the United Kingdom in the period under review 1913–2012.

For example, we will examine how English developed into a global language after World War Two (Brutt-Griffler 2002, Crystal 1997 and Graddol 1997, 2006). This was to promote a shift in focus in examining English at Cambridge away from a number of earlier tasks which had a distinctly academic stance and treated English principally as an object of study (e.g. translation, literature and grammar), to tasks which reflected a social view of English as a means of international communication.

This globalisation of English was not universally acclaimed and a number of writers saw it as a result of ulterior economic and political motives on the part of government, and expressed reservations concerning the limited cultural, political or social awareness in the way it was reported (see Pennycook 1994:Chapter 5, Phillipson 1992:Chapter 8). Whilst sometimes one-sided, these views nevertheless encourage us to think carefully about our tendency to attribute causality solely to individuals and ideas in developments in language testing and teaching.

It is clear that the British Council and Cambridge shared a mutual interest in propagating the English language. By virtue of its global mission in spreading English language and culture around the world, the British Council was able to acquire considerable knowledge and expertise in teaching English as a Foreign Language and for over 50 years (1941–1993), it would be an informed partner in the development of Cambridge English language tests and their availability overseas. The brief of the *Joint Committee of the British Council and Cambridge Local Examinations Syndicate*, established in 1941, was *inter*

alia to collaborate in the actual development and conduct of UCLES EFL exams, make decisions on policy and regulations, prepare examination syllabuses and the general plan of the examination, and cooperate in publicity and finance. The knowledge, professional expertise in English language pedagogy, and the standing of the high ranking academics who served as British Council representatives on the Joint Committee would make a significant contribution to the development of Cambridge English examinations. The Joint Committee would continue until 1993 and was testimony to a close and mutually beneficial working relationship between the British Council and Cambridge in promoting the spread of English and Cambridge English examinations abroad. We will make reference to the influence and work of the powerful Joint Committee throughout this volume.

As we noted earlier progress towards a European Economic Community from the 1970s onwards, with its focus on standardisation across national boundaries, was accompanied by a felt need on the part of governmental agencies to define language teaching and learning goals more precisely at *different levels of proficiency*. This was also to have a marked effect on the work of examination boards in the United Kingdom not least in the way they had to conceptualise language constructs in a more granular fashion, in order to accommodate multilevel tests, than was necessary for tests in the United States which for the most part targeted examinations at a particular level of proficiency.

The influence of language teaching on language testing

A second critical influence on the development of Cambridge English language examinations was the teaching that took place in the English as a Foreign Language classroom and in each chapter we investigate the connection between prevailing approaches to testing and contemporaneous pedagogical practice.

Saville (2009) observes that a close relationship between assessment and the content of learning was an important consideration when UCLES started to administer English language examinations for actual or intending teachers of English in 1913. It was felt that assessment should be relevant to their teaching/learning contexts as well as addressing societal demands for accountability, including maintenance of standards of achievement and impartiality. The assumption that public examinations can help define a teaching syllabus and help to determine learning objectives was to remain a foundational principle in the subsequent development and revision of the Cambridge English examinations right up to the present day.

As far back as the 1920s Jack Roach (Assistant Secretary 1925–45), an influential figure in the early history of Cambridge English examinations, was arguing (against others in UCLES it must be said) for the involvement

of teachers in informing the design and development of Cambridge examinations. Roach felt that the voting power of teachers was negligible and that, as experts in educational needs, teachers should have more influence:

> In the Examining Body of the future the teachers must have a permanent pride of place. . . but the net must be flung wider yet to include professional bodies (who can speak for the country's professional needs), the man in the street, the parents. . . (Roach 1929 'Memorandum of Reform', JOR 1/1d ii).

During his European visits in the late 1930s, as well as gathering political information to be passed on to the French and British governments, Roach promoted CPE and the teaching of English, and he discussed issues in relation to CPE with teachers in UCLES centres on the continent. In this way, he initiated the idea of co-operation with teachers on English language examinations that would continue over the decades (see Hawkey and Milanovic 2013 for comprehensive detail of these relationships). The years that followed saw many further developments in the relations between UCLES and teachers. For example, the minutes of the 1947 meeting of the main committee for the management of Cambridge English Examinations, the *Joint Committee of the British Council and Cambridge Local Examinations Syndicate*, reported that suggestions for modifications of the CPE syllabus would be discussed at a conference with teachers (the annual British Council-Cambridge conference held in London in the post-war years); the representatives of the Syndicate who visited centres in Holland, Belgium and France in 1952 to discuss question papers and syllabuses with teachers reported that 'much useful knowledge was exchanged and some proposals were made for the modification of syllabuses' (Joint Committee Minutes May 1952). In 1971, visits to centres abroad (in most cases individual overseas schools) were seen as valuable by the Syndicate's officers because of 'the opportunities they gave for meetings with teachers. . .' (Joint Committee Minutes July 1971). The minutes of the Joint Committee indicate that they took these opinions seriously. This concern for the voices of a wide range of stakeholders to be heard would be a permanent feature of Cambridge English test innovation and change as can be seen in the accounts of the major developments reported below and in the following chapters.

The impact of developments in approaches to language teaching upon testing at Cambridge will be a major theme in our 100-year survey. We will examine the salient conceptualisations of language constructs with reference to the development of language pedagogy in the United Kingdom over the period 1913 to 2012 and compare these with the prevailing methods employed in English language testing by Cambridge over the same period. Changing priorities in the methods and content of language

9

teaching obtaining at various stages over this period in the United Kingdom included: the Grammar-Translation or Traditional Method, well established in 1913 as the system of Modern Foreign Language education in schools based upon the method used for the teaching of classical languages (with origins in the work of Fick 1793); the direct method promoted in continental Europe for the formal education system, not necessarily involving exclusive use of the target language (Passy 1899); the oral method, Harold Palmer's attempt to systematise direct method teaching procedures and align them with emerging ideas on structural and lexical progression (Palmer 1921a, 1921b); improved reading materials with graded texts (West 1926a, 1926b); standard wordlists for pedagogical purposes (Faucett, Palmer, Thorndike and West 1936, West 1953); the audio-lingual method (Brooks 1960, Fries 1945); the situational approach (Billows 1961, Hornby 1950); and finally the communicative approach (Brumfit and Johnson 1979, Candlin 1986, Morrow 1977, Wilkins 1976, Widdowson 1978) with its focus on the needs of learners to use language for real-life communication accompanied by a sub-skills approach to teaching the four macro skills (Munby 1978).

The consequential aspect of language testing

The area of test impact will not loom large in our history as it only really came into prominence in the 21st century. Prior to the late 20th century there is little evidence of any attention being paid in Cambridge to the macro issues of social impact and test use, the consequential aspects of test validity. Nor to be fair, is there much evidence of any such concern in the wider testing community prior to Messick's (1989) seminal publication on validity. It was not until the 1990s that it came onto the radar of most language testers (Alderson and Wall 1996, Bachman and Palmer 1996, Shohamy 2001, Wall 1997, Wall and Alderson 1993). Milanovic and Saville (1996) appears to be the earliest attempt at Cambridge to address the wider impact issues of Cambridge English examinations. There was, however, an interest in the washback on teaching and learning of its English language tests (impact but at the micro level) from the very beginning in 1913 (see Green 2007).

Spolsky (2004:305) describes how:

> from its beginning UCLES accepted the key role to be played in test development by the 'stakeholders', in particular those schools in various countries of the world that wished to establish examination centres, mainly for their own students. From the earliest years, the Cambridge test writers and their various committees saw themselves as sharing with the schools not so much an examination as the culmination of a teaching process. Before the word 'backwash' had been coined, they regularly asked whether modifications being proposed in the form of the examination would be accepted by the schools.