

Contents

<i>General Editors' Preface</i>	x
<i>Acknowledgements</i>	xi
<i>Abbreviations</i>	xii
Introduction	1
Part 1 Testing as Validity	3
1 Language Testing Past and Present	5
1.1 The Cambridge Proficiency Examination 1913–1945: 'The Garden of Eden', 'the pre-scientific era'	5
1.2 Developments in the 1960s: the move towards a language-based examination	7
1.3 The 1975 and 1984 revisions: 'The Promised Land?'	8
2 The Nature of Test Validity	11
3 Before the Test Event: <i>A Priori</i> Validity Evidence	17
3.1 Theory-based validity	17
3.2 Context validity	19
4 After the Test Event: <i>A Posteriori</i> Validity Evidence	22
4.1 Scoring validity	22
4.2 Criterion-related validity	35
4.3 Consequential validity	37
Part 2 New Frameworks for Developing and Validating Tests of Reading, Listening, Speaking and Writing	41
Introduction	43
5 Test Takers	51
5.1 Physical/physiological characteristics: making accommodations	52
5.2 Psychological characteristics: affective schemata	53
5.3 Experiential characteristics: familiarity	54
6 Context Validity in Action	56
6.1 Task setting	57
6.2 Task demands	68
6.3 Setting and test administration	82

7	Theory-based Validity in Action	85
7.1	Reading	87
7.2	Listening	95
7.3	Speaking	102
7.4	Writing	108
8	Response Formats	119
8.1	Techniques for testing reading comprehension	119
8.2	Techniques for testing listening comprehension	132
8.3	Techniques for testing speaking	143
8.4	Techniques for testing written production	161
9	Scoring Validity in Action	177
9.1	Scoring written production	179
9.2	Scoring speaking tests	191
9.3	Internal reliability of receptive tests	201
9.4	Scores, grading and post-exam validation procedures	205
10	External Validities in Action	207
10.1	Criterion-related validity	207
10.2	Consequential validity	210
	Part 3 Generating Validity Evidence	217
	Introduction	219
11	Research Methodologies for Exploring the Validity of a Test	221
11.1	An introductory note on research	221
11.2	<i>A priori</i> validation: investigating the specification of the construct and the operationalization of the test	222
11.3	Establishing context validity	224
11.4	Establishing theory-based validity evidence	233
11.5	Establishing scoring validity evidence	247
11.6	Establishing evidence on <i>a posteriori</i> validities	259
	Part 4 Further Resources in Language Testing	271
12	Key Sources	273
12.1	Books	273
12.2	Journals	274
12.3	Professional associations	276
12.4	Principal testing conferences	276
12.5	Email lists and bulletin boards	277
12.6	Internet sites	277

12.7 Databases	280
12.8 Statistical packages	280
Postscript	283
<i>References</i>	285
<i>Index</i>	299

1

Language Testing Past and Present

Language tests from the distant past to the present are important historical documents. They can help inform us about attitudes to language, language testing and language teaching when little alternative evidence of what went on in the bygone language classroom remains. Seeing where we have come from also helps us better understand where we are today. The Cambridge ESOL Certificate of Proficiency in English (CPE) has by far the longest track record of any serious EFL examination still in existence, so it is a particularly useful vehicle for researching where we have come from in European approaches to language teaching and testing over the last century. We will trace some significant events in its history to exemplify the developments in the field during that period (see Weir 2003 for a full history of the CPE).

1.1 The Cambridge Proficiency Examination 1913–1945: ‘The Garden of Eden’, ‘the pre-scientific era’

Weir (2003: 2) describes how Cambridge’s formal entry into testing the English of foreign students took place in 1913, when it first offered the Certificate of Proficiency in English (CPE). The examination was based on the traditional, essay-based, native-speaker language syllabus including an English literature paper, the same as that sat by native speakers for university matriculation, and an essay, but also a compulsory phonetics paper, a grammar section and translation from and into French and German. These were complemented by an oral component with dictation, reading aloud and conversation.

The emphasis in this early pre-scientific era was thus on language use, though some attention was paid to form in the grammar and phonetics sections. The ‘scientific’ issue of test reliability was still relatively little understood, at least outside the United States (see Spolsky 1995) and the notion of the ‘connoisseurship’ of an elite group of examiners prevailed. All was thought to be well in this testing Garden of Eden.

1913 CPE Examination

(i) Written:	
(a) Translation from English into French or German	2 hours
(b) Translation from French or German into English, and questions on English Grammar	2 ½ hours
(c) English Essay	2 hours
(d) English Literature (The paper on English Language and Literature [Group A, Subject 1] in the Higher Local Examination)	3 hours
(e) English Phonetics	1½ hours
(ii) Oral:	
Dictation	½ hour
Reading and Conversation	½ hour

The 1913 test corresponded closely to the contents of Sweet's (1899) *The Practical Study of Languages: A Guide for Teachers and Learners* (see Howatt 1984 for details) and mirrored a concern with pronunciation as well as translation. Phonetics occupied a central position in the field of linguistics and language studies which was to survive until the 1960s in tests such as the English Language Battery Version A (ELBA) and the English Proficiency Test Battery (EPTB) used in university admissions (see Davies 2005 for a detailed account of these exams) and even later in the Professional and Linguistic Assessments Board (PLAB) test for overseas doctors wishing to practise in Britain. Grammar translation as a basis for testing proficiency was also to endure into the 1970s in most foreign language testing in the UK and still lingers on in the university sector. In contrast, the testing of English as a foreign language was to progress more quickly.

It is also interesting to note that an oral test (reading aloud and conversation) with associated dictation, was present in an international English as a Foreign Language (EFL) test at such an early stage. This multi-componential approach with a variety of discrete point, integrative and communicative tasks was to differentiate the Cambridge main suite examinations from most of its competitors through the twentieth century. It marks a British/European preoccupation with the *what* we are testing, as against an American preference for the method, the *how* of testing. This contrast was to last throughout the twentieth century until the Test of English as a Foreign Language (TOEFL) Next Generation programme.

Weir (2003: 14) points out how the approach in the first half of the century was to aim for construct validity and work on reliability, 'rather than through the single-minded pursuit of objectivity seriously curtail what CPE would be able to measure. A valid test that might not present perfect psychometric qualities was preferred to an objective test which though

Concept 1.1 Reliability and validity: competing paradigms in test development?

In these early days of language testing, reliability and validity were often seen as dichotomous concepts, a question of where priorities were to be placed. The cardinal guiding principle for Cambridge was construct validity, i.e., appropriateness in what was being measured, followed closely by utility for the teaching community. This does not mean they did not seek to achieve reliability, i.e., consistency of measurement, but reliability was not the overriding determinant of what went into the examination. According to Spolsky (1995), until the work of Roach in the 1940s on improving rater reliability, they appear to have remained relatively immune to psychometric influences from across the Atlantic.

always reliable might not measure that much of value, e.g., not test speaking or writing.'

In America the reverse was true and some aspects of validity were sometimes sacrificed in the pursuit of reliability. It is only with the recent development of TOEFL Next Generation that an attempt has been made to redress the situation by focusing on test activities more relevant to the demands of real-life academic study. Similarly in mainstream education in the USA, there is now increasing public concern over several aspects of validity of a number of the standardized tests that proliferate in school assessment despite their undoubted claims to reliability, i.e., measurement consistency.

We return to the issues of validity in Chapters 2, 3 and 4.

1.2 Developments in the 1960s: the move towards a language-based examination

Concept 1.2 Language tests should only test language

In the early 1960s we see the beginnings of a critical shift in the language testing tradition in Britain towards a view that language might be divorced from testing literary or cultural knowledge. It is thus possible in this period to date the start of a gradual but critical change of the English language examination to one which focuses on language as against an assortment of language, literature and culture. (Weir 2003: 17–18)

Up to this point, the case for a language-based test had been hampered by the desire of linguists to gain academic respectability and recognition for language degree programmes in the older universities by injecting a heavy dose of literature and culture into their courses and examinations.

Weir (2003:19) describes how:

candidates still have to take two other papers in addition to the compulsory 'English Language' paper. However, unlike the previous major revision in 1953, candidates can choose both 'Use of English' and 'Translation from and into English' as two additional papers, which means they do not have to take anything from (b) 'English Literature' or its alternatives.

1966

Oral: Dictation, Reading and Conversation

Written: Candidates must offer (a) English Language and **two** other papers chosen from (b), (c) or (d). No candidate may offer more than one of the alternatives in (b).

- | | |
|---|-----------|
| (a) English Language (composition and a passage or passages of English with language questions. The choice of subjects set for composition will include some for candidates who are specially interested in commerce) | (3 hours) |
| (b) Either English Literature | (3 hours) |
| Or Science Texts | |
| Or British Life and Institutions | |
| Or Survey of Industry and Commerce | |
| (c) Use of English | (3 hours) |
| (d) Translation from and into English | (3 hours) |

In section (b) of the Use of English paper 3 option, multiple-choice items are introduced. This marks a growing interest in improving the reliability of the test overall, at least in terms of the internal consistency of the discrete item components (see Chapter 9). The more consistent the items were with each other in terms of how candidates performed on them, the higher this internal reliability. Spolsky (1978), in line with wider developments in the fields of statistics and linguistics, labelled this the 'psychometric-structuralist' era and Morrow (1979) 'The Vale of Tears'. The latter title was a reaction to an obsessive pursuit of objectivity, not just in tests of micro-linguistic knowledge (e.g., vocabulary) but also, for example, in the Multiple-Choice Question (MCQ) structure and written expression section in TOEFL. This indirect measure was used as an estimate of academic writing ability until the introduction of the bolt-on Test of Written English (TWE) paper in response to consumer wishes in 1986. Breaking language down into its elements also fitted well with the immediate constituent analysis of sentences in vogue with linguists in this period.

1.3 The 1975 and 1984 revisions: 'The Promised Land'?

The 1975 revisions saw the CPE examination taking a shape that, in its broad outline, is familiar to the Cambridge candidate of today and largely

represents the content coverage of language tests at this level across the world. Weir (2003: 24) describes how

the new CPE listening, reading and speaking tests in particular represented major developments on the 1966 revision and echoed the burgeoning interest in communicative language teaching in the 1970s; an increasing concern with language in use as against language as a system for study The 1970s saw a change from teaching language as a system to teaching it as a means of communication as set out and discussed in Widdowson (1978).

In the UK it was reflected in the teaching and publications emerging from CALS at the University of Reading under the influence of Ron White, Don Porter, Keith Morrow and Keith Johnson, and at Lancaster University under the influence of Chris Candlin, Michael Breen and colleagues.

The increased reliance on multiple-choice formats (in papers 2–4) acknowledged the attention international examinations felt they must pay to the demands of objectivity. The concern to improve marker reliability, particularly from the 1980s onwards, also aimed to improve the dependability of the scores in productive tests (papers 1 and 5).

The direct connection between the exam and British culture was completely broken and this potential source of test bias much reduced.

Content of the 1975 Certificate of Proficiency in English

PAPER 1 Composition	(3 hours)
PAPER 2 Reading Comprehension	(1¼ hours)
PAPER 3 Use of English	(3 hours)
PAPER 4 Listening Comprehension	(30 minutes)
PAPER 5 Interview	(approx. 12 minutes)

Weir (2003: 26) describes how

the five papers have replaced the old division of Oral and Written and indicate some movement to recognizing further the need to address the notion that language proficiency is not unitary but partially divisible. It was to take a number of American applied linguists rather longer to discard their firmly held convictions that language proficiency was unitary and that therefore it mattered little what was tested as long as it was done reliably (see Oller 1979).

During the 1980s and 1990s there was, however, a degree of convergence of views on testing internationally, helped in no small part by the growing influence of the Language Testing Research Colloquium, which annually

brought together researchers and scholars interested in language testing from around the world. The birth of the journal *Language Testing* as a result of a weekend meeting of a small group of British testers at Lancaster University in 1980 (see Alderson and Hughes (eds.) 1981) was to promote further the exchange of views across the Atlantic. The advent of the Language Testing list-serve, a web-based discussion forum in the 1990s, similarly promoted the exchange of views and an understanding of different traditions. The growing acceptance, or at least recognition, of international standards for language testing spawned by the drawing up of the American Educational Research Association *et al.* (1974, 1985, 1999) standards made an equally positive contribution. Full details of links to all of these can be found in Part 4.

Now that the channels of communication are open and earlier entrenched positions have softened, the future development of the field will depend on clarifying, codifying and disseminating a framework for test development, administration and analysis that all test developers can buy into. The rest of this book explores what might go into such a framework.

Further reading

Spolsky (1995) is an impressive, scholarly history of the development of ESOL examinations in the USA and Britain, if somewhat predisposed to the psychometric orientation of the former.

Weir and Milanovic (eds.) (2003) gives a full history of the development over a century of a major international ESOL examination the CPE looked at from a British perspective with its humanistic/sociolinguistic leanings.

Index

- authenticity, 20, 32, 56, 61,
73, 101–2, 113, 125, 131,
137, 148–9, 154
- background knowledge, 55, 59, 75–7, 80,
91, 93, 97, 104, 111, 172, 210
- Certificate of Proficiency in English
(CPE), 5–10
 - CPE 1913–1945, 5–7
 - CPE in the 1960s, 7–8
 - CPE: The 1975 and 1984
Revisions, 8–10
- cognitive processing, 20–1, 58, 61–3, 74,
83, 86–7, 108, 110–11, 137, 209,
226, 233, 235–6, 240
- context, 14, 17, 19–21, 26, 35, 37, 48,
51, 55–6, 62, 69, 75, 82–3, 85–7,
93, 96–102, 104–12, 115–17,
119–21, 123, 125, 131–2, 135,
137, 139, 143–4, 149, 152–6, 160–5,
167, 170, 172, 175–7, 181, 187–8,
190–1, 206, 208–9, 211–13, 213–19,
224–6, 228, 231, 244, 250–2, 259,
266, 268–70, 283, 285
- data analysis
 - analysis of variance (ANOVA),
17, 257
 - correlations, 25, 29–32, 34–6, 201, 203,
207–8, 232, 242, 257
 - criterion-related decision consistency,
205
 - factor analysis, 242, 257–8, 270
 - item analysis, 121, 202
 - item-response theory, 26–7, 35,
208, 273
 - observation, 20, 104, 252–7, 268–70
 - profiling, 39
 - protocol analysis, 20, 233, 246,
247, 251–2, 270
 - Rasch analysis, 199
 - multi-faceted Rasch, 35, 199–200,
242
 - reliability estimates, 24, 29–35, 207
 - internal consistency estimates, 22,
31–2, 202–6
 - marker reliability, 9, 15, 35, 48, 123,
199–201, 236, 248, 257
 - standard error of measurement,
33–4, 204
 - t*-test, 234, 241, 246, 251, 257
- ethics, 1, 38–40, 222
- formative assessment, 38–9, 193
- methodologies for establishing validity
 - evidence
 - checklist, 20, 224, 228–30, 236, 249,
251–2, 266
 - discourse analysis, 197, 232,
234, 270
 - document inspection, 222–3,
225, 268
 - expert judgement, 214, 224, 227–30,
249, 251, 252–6, 270
 - interview, 54, 71–2, 105, 141, 151,
153–4, 245–6, 268, 270
 - literature review, 222–3, 225
 - needs analysis, 69, 73–4, 175, 224
 - qualitative procedures, 15, 36,
137, 179, 197, 211, 215,
221, 227, 231–3, 242–6,
251–2, 268, 270, 276
 - verbal report(s), 244–6
 - questionnaires
 - expert judgement questionnaire,
26, 221, 224–5, 228–30,
233–4, 236–7, 240–1
 - students' questionnaire, 224–6,
237–9, 245–6, 259–66
 - research in general, 221–4
 - triangulation, 215, 245
- practicality, 49, 68, 78, 146,
191, 197
- pre-testing, 27, 206, 208

- reliability
- external reliabilities
 - alternate forms (parallel/equivalent), 25–6, 205, 208, 250–8
 - test–retest, 25
 - internal reliability of receptive tests
 - internal consistency, 29–33, 202–4
 - marker reliability
 - scoring production, 34–5
 - scoring written production, 179–91
 - scoring spoken production, 191–8
 - rating procedures, 192–3
 - interlocutor framework, 28, 80, 155–6
 - rater training, 190
 - rating
 - criteria/rating scale, 117, 181–6, 193–7
 - analytical scoring, 183–91, 195–6
 - holistic scoring, 181–3, 188–9, 193–5
 - inter-rater reliability, 34–5, 188, 248
 - intra-rater reliability, 34, 248
 - moderation, 199
 - rating conditions, 200
 - raters, 179–80, 187–8, 190, 192, 199–201, 224, 232
 - standardization, 198
 - statistical analysis: *see* data analysis
- resources for language testing
- books, 273–4
 - data bases, 280
 - email lists and bulletin boards, 277
 - internet sites, 277–80
 - journals, 274–6
 - professional associations, 276
 - principal testing conferences, 276–7
 - statistical packages, 280–1
- response formats
- techniques for testing *listening*
 - comprehension
 - indirect test: matching responses, 132–4; dictation, 134–7
 - information transfer, 141–3
 - integrated test, 101
 - SAQ, 138–41
 - techniques for testing *reading*
 - comprehension
 - indirect tasks: gap filling, 120–3
 - information transfer, 126–31
 - SAQ, 124–6
 - techniques for testing *speaking*
 - indirect tests, 103–4
 - integrated tasks, 154, 157–9, 160
 - mini-situations, 144–6
 - monologic tasks information transfer, 146–9
 - monologic tasks: verbal prompts, 157–61
 - student–student interaction, 149–53
 - co-construction of discourse, 72, 107–8, 145, 153, 155–6, 160, 186, 197
 - reciprocity, 71–2, 146, 149, 152, 156, 230
 - student–examiner interaction, 72, 153–7
 - techniques for testing *written*
 - production
 - essay
 - controlled essay, 166–7
 - open-ended essay, 163–5
 - indirect tasks, 8, 114–17, 175
 - gap filling, 161–3
 - integrated tasks, 77, 87–8, 166–76, 173–5
 - portfolio 66, 111–15
 - rubric, 20, 28, 56–9, 93, 172, 175, 240
- specifications, 25, 54, 203, 222–4
- test taker characteristics, 51–5
- experiential (familiarity), 54–5
 - physical/physiological (accommodations), 52–3
 - psychological (affective schemata) 53–4
- validation, 1–2, 11–12, 14–18, 21, 36–37, 40, 43, 47, 48, 100, 117, 196, 207, 211, 214, 219, 232, 234–5, 244, 251–2, 270, 283
- validation frameworks
- listening, 45
 - reading, 44
 - speaking, 46
 - writing, 47

- validity
 - nature of: a priori validity evidence, 17–21, 43; *a posteriori* validity evidence, 17, 21, 22–40, 43, 259–70
 - consequential validity, 1, 37–40
 - differential validity, 51, 210–11, 260–4
 - impact on individual within society, 214
 - washback in classroom/workplace, 37–8, 211–13, 265–9
 - construct validity, 6, 11–14, 17, 21, 32, 37, 85, 99, 171
 - context validity, 19–21, 56–84, 224–33
 - task-setting, 57
 - known criteria, 63
 - order of items, 64–5
 - purpose, 58–61
 - response format, 62–3
 - time constraints, 65–8
 - weighting, 63–4
 - task demands, 68
 - linguistic (input and output)
 - channel of communication, 72–3
 - content knowledge, 75
 - discourse mode, 68–72
 - functional, 78
 - length, 73–4
 - lexical, 77
 - nature of information, 74
 - structural, 78
 - interlocutor, 79–80
 - speech rate, 80–1
 - variety of accent, 81
 - acquaintanceship, 81
 - number, 81–2
 - gender, 82
 - setting/test administration, 82–4
 - physical conditions, 83
 - uniformity of administration, 83–4
 - security, 83
 - criterion-related validity, 35, 259
 - comparison with different versions of the same test, 208
 - comparison with other tests/measurements, 207–8
 - benchmarking, 36, 107, 209–10
 - comparison with the same test administered on different occasions, 208
 - concurrent/predictive, 35–6, 207, 209, 215, 259
 - scoring validity, 22–35, 177–206, 247–58
 - theory-based validity, 17–19, 85–7, 233–47
 - listening, 95–102
 - reading, 87–95
 - speaking, 102–8
 - writing, 108–18, 237–9
 - scores, grading and post-exam validation procedures, 205–6