contents

series preface  ix
acknowledgements  xii
list of tables and figures  xiii
list of contributors  xvi

introduction  1
   barry o'sullivan

1. test development and validation  13
   barry o'sullivan and cyril j. weir

2. describing language levels  33
   brian north

3. testing linguistic relativity  60
   alan davies

4. componentiality in L2 listening  71
   paul joyce

5. benchmarking and standards in language tests  94
   elif kantarcioğlu and spiros papageorgiou

6. producing an index of word difficulty through learner
   self-assessment data: an application of rasch modelling  111
   toshihiko shiotsu

7. an alternative approach to rating scale development  151
   abdul halim abdul raof

8. quantifying conversational styles in group
   oral test discourse  164
   barry o'sullivan and fumiyo nakatsuhabara
9. a case of testing L2 English reading for class level placement  186
   Anthony Green

10. a model for the development and application of a diagnostic test for trainee foreign language teachers  208
    Deirdre Burrell, Suzanne Graham, Dominique Medley, Brian Richards and Jon Roberts

11. developing affordable “local” tests: the EXAVER project  228
    Adriana Abad Florescano, Barry O’Sullivan, Carmen Sanchez Chavez, David Ewing Ryan, Esperanza Zamora Lara, Luis Alejandro Santana Martinez, Maria Isabel Gonzalez Macias, Mary Maxwell Hart, Patricia Evelyn Grounds, Patricia Reidy Ryan, Roger Anthony Dunne and Teresa de Jesús Romero Barradas

12. the development and validation of an online rater training and marking system: promises and pitfalls  244
    Annie Brown and Paul Jaquith

13. uses and impact of test scores in university admissions processes: the language test as the “hard” criterion  262
    Pauline Rea-Dickins, Richard Kiely and Guoxing Yu

14. assessment in the learning system  282
    John O’Dwyer

Index  305
In the nineteenth century Galton was already using the “new” science of psychometrics, which he defined as the “the art of imposing measurement and number on the operations of the mind” (Galton, 1879: 1). Following Galton, people such as Edgeworth (1888, 1890) began to introduce ideas of scientific enquiry to the study of essay marking in higher education. These early scholars were not to have the sort of impact on assessment in the UK we might have imagined, and by the time the Cambridge Proficiency in English examination, the first formal international language test, was administered in the UK in 1913 the focus had moved from measurement to content, where it was to remain for most of the rest of the century.

However, the situation in the United States was quite different. Here, techniques were devised which offered a solution to dealing with the sheer number of people to be examined across that country. Examples of such techniques include the standardisation of tests of handwriting (Thorndike, 1910) and of composition marking (Hillegas, 1912), which led to the first formal standardised tests (Courtis, 1914). These developments paved the way for a very different approach from that adopted in the UK. The development of the final part of the puzzle, the multiple choice format (Kelly, 1915), meant that very large numbers of people could now be tested across the country, with no compromise on test accuracy. The “modern” tests which emerged from the standardisation movement were perceived to be more scientifically based than the more subjective tests (such as essays or translation into and out of L1) that preceded them (and which remained strong throughout the twentieth century in the UK), and as such their strength was seen to be derived from their accuracy and consistency.
Over the coming half century, the “science” of measurement grew in influence, particularly in the USA, while the “art” of assessment held sway in the UK.

**more recent events**

The changes that have occurred in language testing over the past three decades have seen the area move from a relatively atheoretical, psychometrically driven (in the United States) and practice-led position to one of growing academic success and reach. A number of significant developments in our understanding of language and language development were driven by language testers in this period, the most obvious example being Bachman’s model of language ability, which, though based on earlier work in SLA by Hymes (1972) and Canale and Swain (1980), has had an impact on both language testing and SLA research. In this period, also, the subject has come to be seen by the mainstream of applied linguistics as representing a distinctive field of study, which represents a true “application” of the research findings of linguistics and applied linguistics alike.

Having gone through a revolutionary change in the late 1980s and early 1990s, with the growth in our understanding of the importance of explicit theoretical values to language test design and delivery (driven most obviously by scholars such as Lyle Bachman, J. D. Brown and Adrian Palmer in the United States and Alan Davies, Don Porter, Arthur Hughes, Cyril Weir and Charles Alderson in the UK), there was a general feeling among testers at the turn of the century that things had begun to stagnate. This is, however, not a uniform opinion, as I have felt for some years now that this perception is actually quite wrong. Instead of stagnating, the area has seen the beginnings of a great change, which I see as the “new” revolution, and as a positive force. The major change I am referring to is the dismantling of the old hegemony in which the two dominating centres of language testing, the United States and the UK, provided the dominating theories and practices. This fragmentation has led to the growth in “local” confidence to place the individual test taker, and with this the test context, at the heart of the language testing agenda. One key impact of this change in focus is the recognition that the contextual parameters of test task delivery (e.g. task purpose, timing, knowledge of how task performance will be assessed) should not simply satisfy the predicted demands of the cognitive processes employed in task performance. Instead, if we are to truly place the test taker at the heart of test development, then both the contextual parameters and the
underlying cognitive processes (i.e. the stages of cognition that underlie performance) and resources (e.g. language competence, background knowledge etc.) must be taken into consideration from the perspective of the test taker. This concept forms the basis of the theory of validation discussed by O’Sullivan and Weir in this book.

Half a century ago, when we were first presented with systematic theories of validity (e.g. Cronbach and Meehl, 1955), developers were typically expected to demonstrate the validity of a test by providing evidence related to any one of what were then seen as three distinct types of validity (content, criterion or construct). Within 30 years, Messick (1975, 1980 1989) and Anastasi (1986) had changed the way we viewed validity, arguing that it was a unitary concept. By this they meant that we needed to move away from the position where a single type of evidence would be sufficient to establish the validity of a test to one in which the developer put together a validation argument, to which evidence of various types contributed. In effect, this change meant that we were moving from a position that saw validity as a dichotomy (a test either had it or didn’t have it) to a position of degree (to what extent is it possible to demonstrate the validity of a test?). The unitary theory of validity remains the dominant view to this day, and is interesting not least because it has tended to bring together the two sides of the testing world, the United States, where standardisation and measurement dominated for many years, and the UK, where test content as a reflection of the underlying construct dominated.

This volume offers, I believe for the first time, a coherent overview of some of the key issues currently dominating the profession, bringing together a large group of authors from across the world who are actively engaged in language testing in different ways, from the practitioner to the academic, evidencing, for example, the issue of “joint problematisation” (Sarangi and Roberts, 1999) and the “thick participation” of practitioners with researchers and with consumers and sponsors. The book highlights a number of central ideas, the interpretation of which is coming to be seen as crucial to modern testing (validation, localisation, professionalisation, reflexivity and politicisation).

The book is located in the present, but its main focus is on the future. Each chapter is designed to offer the reader an opportunity to glimpse how language testing researchers and practitioners envisage the discipline developing in the coming years. To achieve the goals set out in this book we need to look beyond present boundaries to embrace the opportunities presented to us by other disciplines and methodologies. This will bring with it a whole new raft of issues and levels of complexity to
testing research and practice, which we will be compelled to deal with in our theories and in the way we conceive, deliver and evaluate assessments. The ongoing contribution of language testing to applied linguistics theory will hinge on our ability to meet these new opportunities.

**current issues**

The themes that surface through the chapters in this book are those that are currently emerging as the central issues that will come to dominate the area of language testing over the coming years. The issues are primarily those of validation, localisation and professionalisation, together with parallel issues of reflexivity and politicisation.

**validation**

Theories of validity have been with us now for over half a century and have moved in a series of stages from a view in which validity was primarily seen as a matter of presenting evidence that the test was accurately testing what it claimed to test, with only a limited expectation of the breadth of evidence required to substantiate any claim. When Messick (1975, 1980, 1989) extended his conceptualisation of validity to include test consequence and the unitary nature of validity (in which a validation argument was constructed from a number of sources of validity evidence), the time seemed right for a major breakthrough from academic theorising (validity) to operational evidence gathering and reporting (validation). However, this failed to happen.

One reason for this situation is the complexity and lack of clarity in Messick's model of validity. The lack of exemplars in Messick's writing has meant that just how specific elements relate to other elements has never been satisfactorily demonstrated, rendering the model at best difficult, and at worst impossible, to operationalise. This is perhaps most clearly exemplified by the recent perception of test consequence as a source of validity evidence (i.e. consequential validity). In fact, test developers should consider the impact on performance of all decisions made during the development process. While undertaking so-called impact studies to look back (typically) on how the test affects, or is affected by, society where there is no direct or indirect effect on performance may soothe our collective conscience, it is unlikely to offer us any meaningful understanding of how that performance should be interpreted. This suggests that test consequence should not be marginalised, as in my view it currently is (i.e. seen as a separate element of the post validity argument), but instead should be recognised as a central aspect of
all test development decisions. In doing so, we acknowledge that consequence drives an ethical and informed test development process by keeping the most important stakeholder (the test taker) at the centre of every decision taken during the process. This is, I believe, how Messick intended consequence to be interpreted.

The reluctance of test developers to commit necessary resources, and the apathy and lack of knowledge among test stakeholders, has meant that there is little or no perceived need for large-scale tests to demonstrate evidence of validity for use in particular contexts. Over the past few decades, however, the atmosphere has begun to change. The push across Europe, in particular (though also seen from Asia to Australia to the Americas and beyond), for transparency and evidence-based proof of the value of test outcomes has stemmed, in my opinion, from the growing and widening professionalism of scholars and practitioners in the field.

**professionalisation**

One major impact of the development of language testing as an area of academic study in its own right over the past 30 years has been the growth of a cadre of professionals with a deepening awareness of theories of language knowledge, use and measurement. One interesting example of this is the Language Testing Forum, an annual meeting of UK testing professionals. When the first group met, in late 1980, there were hardly more than half a dozen participants. The 30th anniversary meeting planned for November 2010 is expected to attract over 40 specialists – a number only limited by the organisers to try to maintain the gathering’s original “feel”. Another interesting aspect of the two events is the profile of the attendees. In the original meeting almost all were British and male; now this is changing, with more women than men, who tend to be younger and more highly academically qualified (either with a doctorate or in the process of gaining one) and from a wider international background.

In the same period, organisations such as the International Language Testing Association (ILTA) and the Association of Language Testers in Europe (ALTE) began to develop professional identities around systematised codes of practices and ethics and offered testers a forum in which theoretical and practical issues could be debated and developed. This has, in turn, led to a more sophisticated understanding of the importance and relevance of test validation. One aspect of this sophistication has been the realisation by testing professionals, particularly in the academy, that, by truly placing the individual at the centre of the
development process, we are recognising the importance of the *local* in the process.

**Localisation**

By localisation, I mean the practice of taking into account those learner-focused factors that can impact on linguistic performance. While this can be seen as an attempt to facilitate the individualising of assessment, it can also be understood as an attempt to take into account characteristics (individual, linguistic, cultural and social) of the learners from a particular population when developing tests for use with that population. Another aspect of localisation is the understanding that by looking primarily to the test taker when developing our tests we are actually taking into consideration aspects of test consequence. This awareness of consequence reinforces the argument presented above, that the current conceptualisation of consequential validity should be rejected in favour of the view that it is an essential aspect of all elements of a validation argument.

Yet another implication of localisation is the recognition of the importance of test context on test development and, by implication, the calling into question of the use of international tests in local contexts without first careful screening (i.e. validation) to establish that they are likely to offer the sort of information required for that context at an acceptable level of accuracy. The example below offers an insight into my argument.

The developers of iBT (Educational Testing Services – ETS) state that:

> TOEFL iBT test scores are interpreted as the ability of the test taker to use and understand English as it is spoken, written, and heard in college and university settings. The proposed uses of TOEFL iBT test scores are to aid in admissions and placement decisions at English-medium institutions of higher education and to support English-language instruction. ETS (2008: 1)

The IELTS Test (designed to allow users to draw inferences on a learner’s ability to cope with the linguistic challenges of academic or non-academic training contexts in UK higher and further education institutions) is described by its developers with the broader claim that:

> IELTS is a secure, valid and reliable test of real-life ability to communicate in English for education, immigration and professional accreditation. IELTS (2010)
Both of these claims must be seen as problematic: the former in assuming that there is no difference in the contexts of English-medium institutions (i.e. where English is the language of communication of the society in which the institution is based, or where another language takes that role); the latter in offering no significant evidence of its suitability for use beyond the original intentions of its developers (academic study or training). Moreover, using either of these tests as a benchmark in a context where learners share a common L1 (which is not English) is clearly problematic, since, without a substantial body of evidence to support this usage, neither test is likely to be demonstrably valid (i.e. for use in that particular context or domain).

Another challenge facing the developers of large-scale international examinations is their accuracy (or internal consistency) in different contexts. Many major examination boards make public internal consistency and standard error of measure estimates based on actual or experimental data. The fact that the ability level of the very large candidature for these examinations is typically very varied means that the reported internal consistency is quite high. When local tests are compared with these figures they often fail to meet the high standards set by their international competitors. However, if we were to make like-for-like comparisons (i.e. the internal consistency figures for the same candidates on both examination papers) it is quite likely that similar estimates would be reported – suggesting that the gulf between the two is probably quite small, if it exists at all. Indeed, the lack of systematic double or multiple marking in some international examinations (such as IELTS speaking and writing papers, for example) means that many local examinations (certainly those employing multiple markers) are likely to be significantly more accurate.

In the past, local examinations were typically seen as being of poorer quality and accuracy than major international examinations. We have seen that this situation is changing as we work together to develop clearer and more transparent validation theories. These theories have highlighted our awareness of the centrality of the test taker and the test context, which, when combined with the spreading professionalism of test developers, has led to the fragmentation of the language testing industry. We are now beginning to understand that tests developed for use in a particular domain or context with a particular population are far more likely to work if the population and domain are taken into account at all stages of design and development.
The book is presented in four sections, each of which is dedicated to a different aspect of assessment. In the first section of this book (Test theories, evidence and our understanding of language) we begin as any book that takes as its focus the future of language testing must begin, in my opinion, with the most current thinking on validation.

The opening chapter of the first section offers a look at the area of test validation by O’Sullivan and Weir. In this chapter, the focus is on the growing influence on test development practice across the world of a set of test validation frameworks, first proposed by Weir (2005). The chapter revisits these frameworks in the light of their continued application, arguing that validation must be placed at the centre of all test development and should not be seen as an optional research issue. One very important aspect of test validity to emerge in recent years is the notion of test level. While the focus in the past tended to be on defining level in relation to other measures of the ability being tested, the development of the Common European Framework of Reference for Languages (commonly referred to as the CEFR) and the positive impact it has had on assessment, particularly, but not exclusively, in Europe, has changed the area dramatically.

In the second chapter of this section, North stresses the significant difference between levels seen as stages of a process, and levels seen as a distinction master/non-master. He moves on to outline the significant recognition the CEFR levels are having in Europe and beyond before discussing the issue of developing descriptors for levels. In this part of his chapter he offers an original and interesting overview of the influence of the different styles/schools of thought on the development of language standards, and sketches out possible development approaches. He also provides an argument for the validity of the CEFR descriptors.

In the third chapter of this section, Davies examines the Linguistic Relativity Principle (also known as the Sapir–Whorf Hypothesis) and seeks to explain how, if at all, it relates to the wider concerns of applied linguistics and the narrower concerns of language testing. This interesting and original piece explores current language testing debates, in particular that of testing English for Specific Purposes (ESP) from an LSP perspective.

The final chapter in the section moves away from the description of test and candidate level to look at a different aspect of test validation, the componentiality of the listening skill. The importance of applying a theoretically sound understanding of the cognitive aspects of a
language skill or ability to the development of a test in that area, stressed by Weir (2005) and Weir and O’Sullivan in Chapter 1 of this book, is revisited here by Joyce. The study reported on in this chapter used a componential research methodology to explore the area of L2 English listening. The findings, that syntactic knowledge is the key component of listening, are discussed here in relation to theories of listening as well as to the practice of test development and the learning and teaching of the skill. This is particularly relevant in that it highlights a limitation of many benchmark descriptors, in which the underlying view of language competence is not supported by a clear theoretical perspective of each language sub-skill.

The second section of the book contains four chapters which look at the application of theory to practice. In the first of these chapters, Kantarcıoğlu and Papageorgiou return to the CEFR and in particular to its use as a benchmark for describing test (and candidate) level. This chapter, which links the earlier chapters by O’Sullivan and Weir and North and reviews the theoretical and practical aspects of standard setting and benchmarking, looking to the authors’ own projects (in Turkey and the UK respectively) for exemplification of the issues involved in engaging in such a process. This paper also explores the interface between the socio-cognitive perspective on validation offered by O’Sullivan and Weir and the process of establishing empirical evidence of a link between a particular test and a language benchmark such as the CEFR.

The second chapter in this section looks at the area of wordlist development. In a departure from the current practice of basing wordlists on analysis of large corpora (see, for example, the work of Nation (e.g. 1990, 2001), Schmitt (e.g. 2000, 2010) and Meara (e.g. 1980, 1996)), Shiotsu employs a sophisticated use of Rasch modelling to develop wordlists from learners’ self-assessment of their own word knowledge. The methodology devised for this project has clear applications for future research in the area.

Abdul Raof then looks to the use of content specialists in the development of what he calls “indigenous” rating scales for assessing conference presentations in the area of civil engineering. The approach taken is founded on earlier quantitative scale development methodologies but offers a unique take on these, and proposes a very practical and user-friendly, yet valid, scale.

In the final chapter in the section, O’Sullivan and Nakatsuhi highlight the need to understand how test takers interact with each other, particularly in terms of how they initiate and ratify topics. This
is of real importance, since it is at these junctures in the interaction that co-construction of discourse (a significant issue in testing speaking using pairs or small groups but also in our broader understanding of spoken interaction) is most likely to be manifested. They outline the development of a set of empirically based measures to quantify conversational styles in groups and then apply these measures to describe conversational styles in group oral tests.

The third section of the book contains four chapters which are devoted to the practice of language testing, all looking at different issues relating to test purpose. The first of these chapters is by Green, who describes the development of the reading component of suite of placement tests for use in English language teaching centres around the world. The constraints implied by the tests’ purpose and candidature and the solutions that emerged during the project are described and discussed, along with the implications for future placement testing projects.

The second chapter in the section looks to the area of diagnostic testing. Burrell, Graham, Medley, Richards and Roberts suggest a model for the development of such an instrument, and exemplify this by describing its application in projects where the resulting instrument was incorporated into a programme of formative assessment in a teacher-training environment. The authors also explore the possibility of using this type of test as a tool for research.

The tendency of international tests to systematically avoid the kind of local traditions and customs in favour of “own-country” culture (e.g. a UK-based international test including elements of UK culture), when combined with the very high costs of these tests, leads to a potentially critical situation for learners in what are often less well off and less “international” communities. Abad Florescano, O’Sullivan, Sanchez Chavez, Ryan, Zamora Lara, Santana Martinez, Gonzalez Macias, Maxwell Hart, Grounds, Reidy Ryan, Dunne and Romero Barradas discuss a suite of tests developed by Universidad Veracruzana in Mexico with the support of a number of UK-based organisations, which marks a first systematic attempt to create a “local”, affordable, and sustainable language test system. The chapter itself is unique in that it was written by the entire team who worked on the development project, and not just by the leaders and/or consultants on that project, thus offering a practical and realistic overview of the process, rather than the filtered narrative such descriptions often present.

In the final chapter in this section Brown and Jaquith return to the general topic of Green’s chapter, that of placement testing. The two papers differ in that, while Green looks at the placement test from a
mainly theoretical perspective, Brown and Jaquith focus on the way in which technology can influence how a large-scale test is managed. In this chapter, they describe the on-line essay marking system that allows them to return test scores for approximately 38,000 candidates in the space of a few days. The chapter first discusses the development of the system before going on to describe the sophisticated rater calibration system put in place, based on multifaceted Rasch analysis of rater performance.

In the first chapter of the final section of the book, Rea-Dickins, Kiely and Yu examine the area of language test use. Using data from a study of a major international test used for language-related university entrance decisions in the UK and other English-speaking countries, they investigate how admissions officers in international programmes interpret test scores in relation to test content and purpose. They also explore the challenges of communication between test theorists, developers and stakeholders, suggesting a need to involve test users, in particular those of high-stakes tests, in the research and development of high-stakes tests.

In the final chapter of the volume, O’Dwyer describes a major curriculum reorganisation in a large pre-university language centre and the various attempts made to develop a formative assessment system to complement the resulting learning system. By learning system, O’Dwyer means the broad triangle of curriculum, delivery and assessment, which together form an integrated whole without which learning is unlikely to be facilitated. The approach taken in the analysis of the situation, namely, to focus on organisational learning and information management, offers a unique and informative perspective that differs from traditional analyses of test washback.

**references**


index

accuracy, 7
Alderson, Charles, 2, 23, 210, 212
American Council on the Teaching of Foreign Languages (ACTFL) scales, 152
Angoff method, 97, 101–2
assessment
see also language testing
art of, 2
criterion-referenced, 35–6
development and benchmarking, 292–5
diagnostic, 210–22
early science of, 1
formative, 287–92, 298–9
in learning system, 282–302
of lexical knowledge, 113–14
quantifying conversational styles, 164–83
rating scale development, 151–61
summative, 298–9
of trainee foreign language teachers, 208–25
understanding levels in, 294–5
variation in, 297–8
Asset Language Project, 35
Association of Language Testers in Europe (ALTE), 5, 16
autonomous competence, 40
Bachman, L., 2, 14–15, 18, 20
Banerjee, J., 264, 265, 270
Basic Interpersonal Communication Skills (BICS), 40
Behaviourally Anchored Rating Scales (BARS), 36
behavioural scales, 104
Behavioural Summary Scales (BSS), 36
benchmarks, 94, 292–5, 300–1
Bilkent University School of English Language (BUSEL), 100–1, 105
BNC frequency, 122
Bookmark method, 97–8
borderline examinee, 100–1
Brennan, R. L., 19
Brown, J. D., 2
BULATS, 25
Business English Certificates (BEC), 25
Cambridge Common Scale project, 50
Cambridge ESOL, 24–5
Cambridge Proficiency in English examination, 1
Canadian Language Benchmarks (CCLB), 95
Canale, M., 40
“Can Do” descriptors, 53
casual speech, 39–40
CEFR, see Common European Framework of Reference for Languages (CEFR)
central executive (CE), 75
Certificate in Advanced English (CAE), 25
Certificate of Proficiency in English (COPE), 103, 284–7, 294–5
Chalhoub-Deville, M., 14–15, 28
Chapelle, C. A., 20
class level placement, testing for, 186–206
cognitive processing, 15
Common Educational Proficiency Assessment (CEPA), 247–60
Common European Framework of Reference for Languages (CEFR),
15–19, 26, 33, 35, 37, 94–5, 294
conversion scale, 44
descriptive scheme of, 51–2
level descriptors, 41–3, 104
reading processes and, 192–4
sub-scales, 41
communicative competence, 27–8, 40
communicative language ability (CLA), modelling, 13–15
competence, 27–8, 40, 53–4
composition marking, 1
comprehension, 190
consequential validity, 4–5, 22–3,
104, 106–7
construct validity, 23
content standards, 96
contextual parameters, 2–3
context validity, 22
conversation, 39–40
group oral, 176–81
quantifying, 165–76
conversational styles, quantifying,
164–83
COPE, see Certificate of Proficiency in English (COPE)
Council of Europe, 95, 106
criterion-referenced (CR) assessment,
35–6
criterion-related validity, 23–4, 104,
105–6, 203
Cronbach, L. J., 3, 19
culture, 69–70
Cummins, J., 40
Cureton, E. E., 19
curriculum innovation, 287, 292–5
curriculum levels, 34–41
cut scores, 95–6, 98, 103
Davies, A., 2
Delandshere, G., 210
descriptors, see language proficiency descriptors
descriptor scales, 43–6
diagnostic assessment, 210–22
EBB scale, 153, 154
Edgeworth, F. Y., 1, 19
Elder, C., 244, 245, 246, 247
English Profile project, 50–1
essay marking, 1
European Association for Language Testing and Assessment (EALTA),
16
European Association for Quality Language Services (EAQUALS), 34
European Language Portfolios (ELPs),
17, 33, 52
EXAVER project, 228–43
aims of, 230–1
context for, 229
evolution of, 234–9
overview of, 231–4
results of, 239–41
First Certificate in English (FCE), 25
Fleming, M., 46
foreign language teachers, diagnostic testing of, 208–25
Foreign Services Institute (FSI), 152
formative assessment, 287–92, 298–9
“four skills” descriptor categories,
42–3
Fulcher, G., 16, 26, 35, 104, 154–5,
159, 161
functional descriptors (Can Dos), 16
Galton, F., 1
Glaser, R, 35–6
goal-orientation, 165, 176
Goffman, E., 54
grammar, 72–3
group oral tests, 164–83
Gumperz, J. J., 64, 65
Hargreaves, P., 37
Henning, G., 113
Herder, 61
Hughes, Arthur, 2
Huhta, A., 114, 210, 211, 212
Hulstijn, J., 40
Humboldt, W. von, 61

iBT, 6
impact studies, 4
index of lexical difficulty, 114–50
Integrative Grammar Test (IGT), 74
interactional contingency, 165–6, 176
International Civil Aviation Organisation (ICAO), 107
International English Language Testing System (IELTS), 6, 67, 263–80
international examinations, accuracy of, 7
International Language Testing Association (ILTA), 5
International Legal English Certificate, 25
international students, 276–9
Item Response Theory (IRT), 97–8, 112–13

JACET 8000, 119–20, 122–3, 127–8
Jefferson, G., 169
joint problematisation, 3

Kanda English Proficiency Test (KEPT), 77–9
Kane, M. T., 20
Knoch, U., 210, 244, 246, 247
Kuehn, P., 73

L2, see second language (L2)
L2 listening
  componentiality in, 71–89
  language testing and, 88–9
  literature review and methodology, 72–9
  metacognitive listening strategies usage, 77
  phonological awareness and, 75–6
  phonological modification knowledge and, 74–5
  research procedure, 71–2
  sentence stress awareness and, 76–7
  study results of, 79–85
  syntactic knowledge and, 72–3
  vocabulary breadth and, 73–4
  working memory and, 75

L2 reading, testing, for class level placement, 186–206
Lado, R., 42
language, thought and, 61–2
language ability
  Bachman’s model of, 2
  development of, 38–41
language acquisition, as unidimensional, 37–8
language audit, 38–9
Language for Specific Purpose (LSP) scales, 36, 42, 67
language ladder, 35
language learning contexts, 127–8
language levels, 34–41
language proficiency
  vs. competence, 53–4
  core, 40–1
  as unitary, 67–8
  via varieties, 67–8
language proficiency descriptors, 33–54
categories of, 41–3
conceptualisation, 47
construction, 47–8
development of, 46–9
level “C”, 40–1
levels, 34–41
style, 43–6
usefulness of, 33–4
validation, 48–54
language structure, language use and, 66–7
language testing
  see also assessment; test development
  benchmarking and standards in, 94–107
  birth of modern, 1–2
  changes in, 2–4
  current issues in, 4–7
  of foreign language teachers, 208–25
  of L2 English reading, for class level placement, 186–206
  L2 listening and, 88–9
  in learning system, 282–302
  LRP and, 65–70
  online marking systems, 244–60
language testing – continued
quantifying conversational styles, 164–83
role of, in university admissions, 262–80
topic identification in, 168–76
validation, 103–7
validity, 68–70
Language Testing Forum, 5
language use
language structure and, 66–7
modelling, 13–15
social theories of, 14
learning system
assessment in, 282–302
expectations in, 283–4
increasing standards in, 297
validity, 299–300
Levinson, S. C., 64, 65
lexical competence, 111–12
assessments of, 113–14
Linguistic Relativity Principle (LRP), 60–70
language testing and, 65–70
linguistics, 2
literacy, 40
localisation, 6–7
local language variations, 69–70
Loevinger, J., 19
LRP, see Linguistic Relativity Principle (LRP)

Many-facet Rasch (MFR) analysis, 247
Maynard, D. W., 169
McNamara, T., 14
meaning, 65
measurement, science of, 2
Meehl, P. E., 3, 19
Messick, S., 4–5, 19–20, 22
Metacognitive Awareness Listening Questionnaire (MALQ), 77
metacognitive listening strategies usage, 77
miscommunication, 61–2
Mislevy, R. J., 20, 50
multi-competence model, 67
multiple choice format, 1
National Assessment of Educational Progress (NAEP), 95–6
No Child Left Behind project, 47
North, B., 152, 159, 161
Oller, John, 67
O’Loughlin, K., 264, 265, 271
online training and marking, 244–60
background to development of, 247–9, 253–8
benefits and efficacy of, 245–7
current developments in, 258–60
early developments in, 249–52
oracy, 40
oral language tests, scoring of, 182–3
organisational learning, 292–5
O’Sullivan, B., 14, 17–18, 25, 26
Palmer, A., 2
performance level descriptors (PLDs), 104
performance standards, 96
phonological awareness, 75–6
phonological modification knowledge, 74–5
phrasal verbs, 112
placement testing, 186–206
Porter, D., 2
Principle Components Analysis (PCA), 82–3, 118
professionalisation, 5–6
prosody, 76
psychometrics, 1
QALSPELL project, 25
quantitative dominance, 166–7
Rasch modelling, 112–13, 118–23, 127–8
raters
online training for, 244–60
reliability, 19
rating scales
alternative development approach for, 151–61
definition of, 151
intuitive method of development, 152–3
rating scales – continued
qualitative method of development, 153
quantitative method of development, 153–5
reading
pace of, 191–4
testing L2, for class level placement, 186–206
reading comprehension, 190
Sapir, E., 62, 63
Sapir-Whorf Hypothesis, see Linguistic Relativity Principle (LRP)
scoring validity, 22, 24, 104, 105
second language (L2)
see also L2 listening
metacognitive listening strategies usage, 77
phonological awareness, 75–6
phonological modification knowledge, 74–5
sentence stress awareness, 76–7
syntactic knowledge, 72–3
sentence stress awareness, 76–7
short-term store (STS), 75
Simulated Oral Proficiency Interview (SOPI), 245
social consequences, 22–3
social theories of language, 14
socio-cognitive framework, 20–6, 186
socio-linguistic ability, 45
Spearman, C., 19
Standard English, 69
standardised tests, 1
standards, 94–107
content, 96
in learning organisation, 300–1
performance, 96
standard setting, 95–6

differences in judges’ perceptions and, 100–1
institutional implications of, 102–3
issues regarding, 99–103
judgemental nature of, 99–100
process of, 96–9
use of multiple methods of, 101–2
validation of, 103–7
storytelling, 40
student teachers, diagnostic testing of, 208–25
Studies in Language Testing series, 25
summative assessment, 298–9
Surveylang project, 25
syntactic knowledge, 72–3
Taylor, D., 27
Taylor, L., 16, 24
test accuracy, 7
test context, 6, 7
test development
see also language testing
for class level placement, 186–206
of diagnostic test for trainee foreign language teachers, 208–25
EXAVER project, 228–43
socio-cognitive framework for, 20–6
validity and, 13–28
test impact, 22
test performance, validity and, 4–5
test reliability, 24
tests, standardised, 1
test specifications, 16, 25–6
test validation, model of, 19–24
test validity, 3, 299–300
see also validity/validation
test impact, 22
theory-based scale, 104
theory-based validity, 104
thought
language and, 61–2
modes of, 65
TOEFL, 77–9
topic identification, 168–76
training systems
face-to-face, 245
online, 244–60
Trim, John, 34, 45
unidimensionality, 37–8
unitary theory of validity, 3, 19–20
university admissions, role of
language tests in, 262–80
University of Veracruz, 229, 239
validity/validation, 3
consequential, 4–5, 22–3, 104, 106–7
construct, 23
context, 22
criterion-related, 23–4, 104, 105–6, 203
of descriptors, 48–54
of language testing, 103–7
learning system, 299–300
Messick’s model of, 4–5
model of, 19–24
questions, 68–70
scoring, 22, 24, 104, 105
socio-cognitive framework for,
20–6, 186
of standard setting process, 103–7
test, 299–300
test development and, 13–28
theories of, 19–20
theory-based, 104
theory of, 3, 4–5

Van Lier, L., 39
vocabulary breadth/size, 111–12,
113–14
washback, 22
Weir, C., 2, 16–17, 20, 24, 103–4, 186,
203
Whorf, B. L., 62–5, 66
Widdowson, H. G., 27–8
word difficulty
frequency and, 123–5
index of, 114–50
relative, 112
word families, 115
word frequencies, 112, 123–5
working memory, 75

Yes/No checklist test, 101–2, 113–14,
128
Young, R., 170

Zayed University, 25