Chapter 2
Research Literature

David Coniam and Peter Falvey

Abstract In this chapter, high-stakes assessment, educational standards and benchmarks are first discussed. These are subsequently elaborated upon and discussed within their theoretical contexts and, particularly, within the context of language assessment. The ways in which assessment paradigms have changed in recent decades are also discussed to highlight their influences on the benchmarking project.

Changing Paradigms

In this chapter, high-stakes assessment, educational standards and the benchmark-setting phenomena are first placed in their theoretical contexts and then within the context of language assessment.

The major testing and assessment paradigm that was used in the last half of the twentieth century stressed the reliability of test items over their validity because of legitimate concerns about consistency and fairness in testing (Moss, 1994). In this paradigm, language tests tended to test segments of language (e.g. slot and gap-filling exercises and multiple-choice items) rather than discourse-based 'chunks' of language above the level of the sentence. The purpose of testing segments of language was to avoid testing elements of language other than the construct or skill being assessed. It was a paradigm that focused more on the act of *testing* than on the more holistic paradigm of *assessment*.

The connotation of the term *assessment* and, in particular, the term *high-stakes assessment* embraces a wider set of parameters than does the term *testing*. Advocates of the testing paradigm would tend to avoid any form of integrated testing. Those who were opposed to integrated forms of testing had legitimate concerns (Lee, 2006).

D. Coniam · P. Falvey (✉)
Department of Curriculum and Instruction, Faculty of Education and Human Development, The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: falvey@eduhk.hk

D. Coniam
e-mail: coniam@eduhk.hk

© Springer Nature Singapore Pte Ltd. 2018

D. Coniam and P. Falvey (eds.), *High-Stakes Testing*,
https://doi.org/10.1007/978-981-10-6358-9_2

Their concerns focused on the psychological distance between starting and ending the task, the potential areas for distraction or being misled during the earlier activities and the problems associated with grading/marking the final product if it had been 'contaminated' by what had occurred earlier (Lee, 2006). It is easy to sympathise with those who objected to integrated tests. Indeed, it was this basic stance with its aim of testing only that which should be tested that led to language being segmented so that the test items that were produced could be seen to test one element of language only.

Reliability-focused testers felt that unless a test is reliable, questions about its validity are not worth considering (Chapelle, 2012). Thus, questions of validity were shrugged off and demoted to secondary status. This perspective allowed for the growth of a form of testing known as *indirect testing*, a form of testing which states that although the indirect test may lack validity, it is possible to infer from the test taker's score how well the skill or knowledge that is the focus of the test has been mastered (Hughes, 2003). Such tests, including the Educational Testing Services' TOEFL dominated the language testing market for years because it could be proven that their tests were reliable (test–retest results were consistent) even though the earlier version of the TOEFL paper-based test (PBT) contained no form of direct testing of communication through speaking (an oral or speaking test) or writing (writing an extended piece of prose) (Chalhoub-Deville & Turner, 2000). The format of the TOEFL further evolved from a paper-based test to a computer-based test, then to an Internet-based test (iBT), in association with developments in the theories motivating test design; the TOEFL iBT pays more attention than did the TOEFL PBT to communicative competence and the ability to use language knowledge in relevant contexts (Taylor & Angelis, 2008).

The University of Cambridge, however, had persevered—since the inception of its Local Examinations Syndicate (UCLES) in the mid-1880s—with written tests and, later, interactive oral tests. In the face of twentieth-century concerns about the reliability of tests, it was believed by UCLES (now known as Cambridge Assessment) and its advocates that there was a place for direct tests which they later balanced with shorter segment tests of language through multiple-choice tests. In answer to criticisms about the lack of reliability on the written and spoken tests, UCLES worked hard to ensure that writing raters were standardised (all raters come to Cambridge to be trained in grading and are standardised) (Milanovic, 2016; Weir, 2005).

In addition, with the use of new technologies such as videos and DVDs, the training and standardisation of oral raters have become much more systematic and reliable. Cost, of course, is a major consideration. The results of multiple-choice tests can be scanned into a computer, and results processed very quickly (Bachman, 1990; Dooey, 2008). Direct tests, on the other hand, require human resources—assessors who have to be trained first. As will be seen below, in the midst of these paradigm changes, large testing organisations such as Educational Testing Services (ETS) have, over the past thirty years, begun to make available tests of spoken and written English to complement their original multiple-choice grammar, reading and listening tests.

Before discussing changes in English language testing, changing paradigms in testing and assessment practices in other places will be discussed. This is because changes in English language testing and assessment often follow innovations that have been made elsewhere (Eraut, 1994; Gipps, 1994). Eraut charts change to testing in the professional world of airline pilots, lawyers and doctors. Gipps (ibid) proposes a form of assessment that the title of her book *Beyond Testing: Towards a Theory of Educational Assessment* encapsulates. She advocates a holistic, constructive approach to assessment which de-emphasises the indirect tests, which deselect so many test takers, in favour of regular, formative assessment rather than summative assessment, and profiles of what students *can* do rather than scores which tell parents and employers little except what the student *cannot* do.

As parents and teachers reacted to old-fashioned methods of reporting the results of tests in a norm-referenced manner, Biggs (1996) illustrated this issue for Hong Kong in discussing how 80% of the school cohort at age 16 (equivalent to US Year 11) are deselected by public examinations. The problem with this form of testing and reporting is that only 20% of the whole cohort is deemed to have satisfied the examiners. The rest, having been deselected in a norm-referenced manner, have no means of showing prospective employers that they do, in fact, possess some academic or vocational qualities. Thus, Gipps (ibid) and Tang and Biggs (1996) advocated all-inclusive reporting of achievements for students, both stating that what should be reported is what students *can* achieve, rather than what they *cannot* achieve. Biggs (2012) reiterates that the use of criterion-referenced assessments better addresses and reflects whether and in what way students have achieved the learning objectives.

Trends in Assessment and Evaluation

Given the background above, two trends have emerged over the past four decades in the area of assessment and evaluation. These are criterion-referenced assessment (often linked to a task-based curriculum and assessment procedures in English language assessment) and competency-based assessment (often linked to vocational, and, increasingly, professional-based training and assessment) (Hudson, 2005).

On the issue of competency-based assessment, Brindley states:

> Competency-based models of vocational education and training have in recent years dominated the educational landscape in Australia, the UK and New Zealand. They have also begun to exert a significant influence in the field of language learning. (1995, pp. 145–164)

Brindley (1995, pp. 1–2) stresses the need for a theoretical approach to assessment and discusses the necessity for test developers to begin with a clear theoretical conceptualisation of the abilities they are assessing and to 'reality-test' their constructs against data from the target language use situation. Brindley's (1998) review of the issues inherent in outcome-based assessment and reporting in language learning programmes warns against the problems of assessing individual progress in language learning, especially when combining formative with summative reporting and in

matters of reliability and validity in outcome statements. He states that these problems can be alleviated by close consultation between policy-makers, administrators and practitioners. He discusses these issues in the context of school assessment and stresses the need for teacher professional development.

In Canada, Citizenship and Immigration Canada (CIC) decided to implement a project to develop language benchmarks for immigrants to Canada. Four TESL Canada Learners' conferences, held in 1994, discussed the issue, and the working document *Canadian Language Benchmarks* was produced in 1996. *Canadian Language Benchmarks* presented two sets of benchmarks—*Canadian Language Benchmarks*: *English as a Second Language for Adults* and *Canadian Language Benchmarks: English as a Second Language for Literacy Learners*. In 2000, the *Canadian Language Benchmarks 2000* was published. The work aimed at making the language benchmarks for Canada a practical and usable document (Pawlikowska-Smith, 2002). It elaborated the theoretical basis of the language benchmarks, providing examples of different language competence components, and demonstrating different levels of language proficiency, from basic language proficiency to full fluency (Fleming, 2015). In 2012, a revised version of *Canadian Language Benchmarks* was published, with an updated theoretical framework (see http://www.cic.gc.ca/english/pdf/pub/language-benchmarks.pdf). The benchmarks were validated against the Common European Framework for Language, the American Council for the Teaching of Foreign Languages (ACTFL) and the Quebec version of the benchmarks. The comparisons showed that the benchmarks were consistent with the theoretical concepts of the language frameworks as well as the key principles underlining other language frameworks. The validations also indicated that the Canadian Language Benchmarks were valid and reliable for a variety of purposes—including high-stakes ones—and in a variety of contexts—including community, workplace and academic (Hajer & Kaskens, 2012).

Other developed countries which possess an academically and professionally trained language teaching workforce (such as Australia) require teachers to undergo professional training before new forms of assessment can be exploited successfully (see Elder, 1994). In the area of assessment of teacher classroom language, Elder found significant correlations between the results of assessments of ESL teachers and subject specialists, indicating that agreement can be reached when assessing teachers' classroom performance. One of the major objectives in making the Hong Kong language benchmark Classroom Language Assessment (CLA) rating scales criterion-referenced was the desire for transparency so that teachers themselves, as well as informed laypersons, could, with training, reach similar grades when viewing videos of English teachers and assessing them on the four CLA scales. The Oral Proficiency Interview (OPI) of the American Council for the Teaching of Foreign Languages (ACTFL) is another example of a criterion-referenced certification test. The test results demonstrate to what extent teachers' levels of language proficiency are sufficient for them to perform teaching duties (Bachman & Purpura, 2007). A further factor in choosing criterion-referenced benchmarks is the positive washback effect which the benchmarks can engender. McDowell (1995) states:

> …… it was felt that candidates for the ELSA would work towards establishing strategies for 'passing' the test by becoming test-wise and teachers would likewise seek ways to prepare their candidates to maximise their chances of success. This has already proved to be the case. (1995, p. 19) (our underlining)

The major effect of this changing paradigm of assessment on schooling is that criteria are specified for the various stages of a student's school life. Instead of being ranked against other students, the student is ranked against a series of 'can do' statements, competing against known, agreed, sets of criteria. These act as standards of achievement which schools can report to parents and, eventually, employers. Hudson (2005) reports further developments in criterion-referenced benchmark assessment, such as the Canadian Language Benchmarks, the Common European Framework and the Assessment of Language Performance Project (Hudson, 2005).

In cases where assessors have been initially trained and standardised against rating scales and descriptors, it is extremely important, for purposes of reliability, that whenever a new batch of assessments is to take place, further training is provided, particularly if there has been a significant time-gap between the initial training and the administration of the new batch of assessments (see Lumley & McNamara, 1995). Assessor training is important to ensure that all assessors assign grades in a consistent way, especially when a test is graded by a group of assessors (Sercu, 2004). However, in Baird et al.'s (2004) research, contradictory findings emerged to the effect that exemplar works and discussion about students' work did not contribute to more reliable marking. Baird et al. (2004) explain that in a well-developed community of assessment practice, one possible result of recent developments in explicit marking schemes does not necessarily need exemplars and discussions to produce accurate marking.

Language and Language Teacher Standards

Sykes and Wilson (1988) report on the work of the National Board for Professional Teaching Standards which investigated the implications of introducing procedures for the voluntary certification of teachers to a standard of 'advanced competence' with advanced levels of knowledge and skill. Foreign language teacher education and certification requirements have changed considerably over the past forty years. Whereas proficiency was not an issue in the era of audio-lingual methodology, with teachers supposedly being able to compensate for their lack of proficiency by taking their students to a language lab, foreign language teacher standards have recently become increasingly important, with more attention being paid to teachers' ability to use language in the classroom (Donato, 2009).

In developing countries, teacher certification is nowadays becoming increasingly important, although not as well established as teacher certification in developed counties (Elder & Kim, 2014; Fischer, 2013; Pearson, Fonseca-Greber, & Foell, 2006). For example, in Indonesia, the government started a national-wide teacher certification programme with the aim of certifying as many as 2.3 million teachers

by 2015, although such a certification process did not focus specifically on language standards (Fahmi, Maulana, & Yusuf, 2011). Within the context of Asian languages, Sadtono (1995) was an early caller for the certification of non-native speakers of ESL. His intentions were broadly similar to those investigated in the LPATE case study reported in this book—although his proposals did not involve the use of criterion-referenced assessments.

A discussion of some of the relevant research conducted with regard to benchmark or certification procedures in the context of second language teachers' language standards will now be presented on a country-by-country basis.

USA

In the USA, a variety of language standards agencies have been set up to guarantee the language standards of language teachers, such as the Interstate New Teacher Assessment and Support Consortium (INTASC), the American Council on the Teaching of Foreign Languages/National Council for Accreditation of Teacher Education (ACTFL/NCATE) and National Board of Professional Teaching Standards (NBPTS) (Donato, 2009).

As early as the 1990s, most USA states had some measure of certification for ESL instructors in place (see, e.g. Grant, 1995; Kornblum & Garschick, 1992; Thomas & Monoson, 1993). Many of the certification tests, however, appeared to focus on *subject-matter knowledge*, rather than on *language ability* per se, although Thomas and Monoson state, in relation to International Teaching Assistants (ITAs), that:

> student complaints to legislators led to 20 states mandating higher educational institutions develop policy on oral English language proficiency of international teaching assistants. (ibid, p. 195)

The language proficiency of ITAs has been proven to be crucial in American classrooms, with research indicating how greater ITA language fluency leads to increased students' perceptions of clarity and credibility (Li, Mazer, & Ju, 2011).

In the USA, many states have made it mandatory for international students to be assessed on their oral language proficiency before they are allowed to teach. The commonly used assessments are Speaking Proficiency English Assessment Kit (SPEAK) test, locally developed teaching simulation tests, or the new TOEFL which has an integrated speaking performance component (Farnsworth, 2013). The TOEFL has been considered to measure the same general speaking proficiency as the locally developed teaching simulation tests, although the teaching simulation tests address more than general English language proficiency. The tests involve candidates giving both a prepared lecturer in a specific content area and a short explanation of course material. The Test of Oral Proficiency is such an example. The teaching simulation tests fall in line with initiatives to assess pedagogical content knowledge—an issue which Carlson (1990, pp. 157–163) discussed as far back as 1990:

> to create items that require something other than a bit of pedagogical knowledge and a bit of content knowledge. (1990, p. 159)

The test designers thus attempted:

> to create test items that require application of pedagogical knowledge to specific content areas. (1990, p. 160)

At the end of the twentieth century, the *No Child Left Behind* act in the USA brought further attention to teacher effectiveness and student achievement (Chambless, 2012). Along with a concomitant increase in the number of immigrant students who were English language learners (ELL), teachers were expected to have the language knowledge to support ELL—attending, for example, to students' oral language development, supporting their academic language, and being sensible to cultural differences (Samson & Collins, 2012). Such a requirement further supported the stance that teachers need pedagogical content knowledge. Across different states in the USA—Massachusetts, New York and Florida for example—there are different teacher certification examinations focusing on the linguistic features of English, and the effective use of language to cater for different student needs (Samson & Collins, 2012).

In addition to assessing language teachers through examinations, across different states there are add-on ELSO (credentialing teachers of English to speakers of other languages) certification programmes. Certain add-on ELSO programmes have, however, been criticised for giving insufficient attention to prospective teachers' linguistic knowledge and abilities to teach linguistic knowledge to students (Reeves, 2010).

Language requirements have also been put in place for teachers of world languages (including a range of languages such as French, Spanish or German). The American Council on Teaching of Foreign Language (ACTFL) developed the OPI and WPT to assess what foreign language teachers need to know, with all language teachers having to meet minimum requirements before being permitted to take up teaching duties (Burke, 2013). The teacher language proficiency tests in the USA expect teachers to understand language acquisition and to create a supportive language classroom. Teachers are expected to use appropriate teacher talk to give instructions, to ask questions, to check learners' understanding and to guide discussions (Pearson et al., 2006).

In Sect. II, discussion will move to how issues such as those which Carlson cites above were dealt with in the development of assessment instruments used to assess the speaking and writing of English language teachers.

Canada

Since 1992, a number of documents related to language policy for teachers have been published, such as the Canadian Language Benchmarks 2000, Canadian Language Benchmarks 2012 and Language Instruction to New Comers to Canada (LINC) curriculum guidelines (Haque & Cray, 2007). These benchmarks guide teaching approaches, assessments, activities and assessments in classrooms.

In Canada, where each of the 13 Canadian provinces has its own policies for education, there are agreed expectations of language teachers. All language teachers are expected to meet three minimum requirements: a post-secondary degree from an accredited university, a qualification from an accredited teacher education programme and evidence that the person will uphold the standards of the teaching profession (Salvatori, 2009). Specifically, most Canadian jurisdictions issue a generic Kindergarten to Year 12 teacher certificate for English as a Second Language (ESL) or French as a Second Language (FSL) teachers, which limits the number of teachers who may be assigned to the teaching profession (Salvatori, 2009).

UK

Alderson et al. (1997) reported on a project with university students of French in the UK to examine their *subject-matter knowledge* in French. The University of Cambridge Local Examinations Syndicate/Royal Society of Arts examined the notion of benchmarks *(language ability and language awareness)* for entry to their Certificate in the Teaching of English as a Foreign Language programmes because of concerns that teachers applying for their courses leading to teacher examinations did not possess minimum levels of subject-matter knowledge/language awareness.

In the UK, since 1998, primary English teachers have been required to demonstrate their subject knowledge through a literacy test when exiting Postgraduate Certificate for Education programmes. Further, the Department of Education has made it a requirement that all in the state sector in the UK pass a literacy test organised by the UK Department of Education (see https://getintoteaching.education.gov.uk/how-to-apply/passing-the-skills-tests). The literacy professional skills test is divided into four sections: spelling, punctuation, grammar and comprehension.

The Teaching Knowledge Test (TKT*)* was introduced by Cambridge ESOL in 2005 for practising or trainee teachers who teach English to Speakers of Other Languages (ESOL). According to Cambridge ESOL, the test focuses on testing knowledge about teaching (Spratt, 2015). The test papers included matching and multiple-choice tasks on three modules:

1. Language and background to language learning and teaching
2. Lesson planning and use of resources for language teaching
3. Managing the teaching and learning resources (see http://www.cambridgeenglish.org/teaching-english/teaching-qualifications/tkt/about-tkt/).

The TKT does not assess candidates' teaching ability, performance in classroom situations or English language proficiency. Rather, it assesses candidates' declarative or received knowledge about English language teaching. The knowledge components assessed in the TKT included subject-matter knowledge, knowledge of the content, pedagogical content knowledge and general pedagogical knowledge (Spratt, 2015).

Australia

A report to the Queensland Board of Teacher Registration (1992) discussed the situation with regard to the teaching of languages other than English in Australia. McKay (1995, 2000) reported on the creation of *Bandscales* in Australia, where scales were trialled on Australian students and profiling attempted. The ESL *Bandscales* are proficiency scales tied to English language development, and they can be applied to different schools and education systems where English is a medium of instruction (McKay, 2000). Arising from the Australian experience, McKay and Ferguson (2000) debate if the *Bandscales* are transferable and in doing so present a set of questions to guide the setting of standards for L2 learners both in Australia and in Mainland China.

In an article reviewing the major contribution of Penny McKay to the field of TESOL [Note 1], Dooley and Moore (2009) noted that the *Bandscales* had three prominent features: first, the *Bandscales* were grounded in actual practice—which can reflect and guide real classrooms. Second, the *Bandscales* were informed by the assessment framework: the *Bandscales* were developed through an iterative process that interrogated researchers' and teachers' classroom observations related to second language acquisition research and related assessment research. Third, Penny McKay put considerable effort into extensive consultation with relevant stakeholders, to ensure the impact of the *Bandscales*.

Japan

In 2003, an action plan on English language education was introduced in Japan. The plan was introduced as a response to the critique that English language education focused too much on exam-related learning and did not meet the multiple needs associated with globalisation. In an attempt to ensure the quality of English language teachers, proposals were put forward for English language teachers to obtain a proficiency level equivalent to certain recognised local or international language tests. Among the proposed recognised tests and suggested levels were: the pretest level of the STEP test (the English certificate examination designed by Japan's Eiken Foundation the former *Society for Testing English Proficiency)*; a score of 550 in the Test of English as a Foreign Language (TOEFL), or 730 or above in the Test of English as an International language (TOEIC) examination (Butler & Iino, 2005).

The EIKEN Test ('Jitsuyo Eigo Gino Kentei', or Test in Practical English Proficiency) is one of the most widely used English language testing programmes in Japan. The test is offered at seven levels: from Grade 1 (the highest level, equivalent to CEFR C1) down to Grade 5 (the lowest level, equivalent to CEFR A1). The EIKEN Test has been used to assess English language teachers' proficiency, with Grade Pre-1 (B2) set as the benchmark level for English teachers (Eiken Tests, 2017).

As of late 2016, Grade Pre-1 or higher on the EIKEN had been achieved by 30.2% of English teachers in middle school and by 57.3% of English teachers in high school. The target set by the Japanese government was that 50% of middle school and 75% of high school English teachers would achieve the Grade Pre-1 proficiency level by 2017 (see Kimura et al. 2017) for a discussion of developing classroom language assessment benchmarks for Japanese teachers of English.

China

The People's Republic of China (PRC) has long struggled to increase the competence and professional skills of its teachers of English language, with expansions in the provision of English language education creating a shortage of qualified English teachers. In 1988, only 30% of English teachers in junior secondary schools and 26% in senior secondary schools met the language requirements laid down by China's State Education Commission—a factor which greatly affected the envisioned goal of English language teaching in China (Hu, 2005). In addition to a considerable increase in the number of teacher education institutions and universities, there was also a substantial increase in the provision of in-service teaching training courses (Hu, 2005).

Despite the increase in provision, a coherent framework targetted specifically at English language teacher proficiency across diverse pre-service and in-service teacher education programmes in China (Hu, 2004) is still lacking. Whereas all English teachers in China are expected to pass the Test for English Major students TEM-4, such a test focuses on general English language proficiency rather than the specific English language proficiency of English teachers cited by Elder (1993).

In 2016, a national social science project at Beijing Foreign Language Studies University (Han & Qu, 2016) sets out to investigate potential ways of assessing English language teachers' language standards in Mainland China, despite the fact, as mentioned, that currently no established language standards exist for English teachers in China.

The Mainland China issue is massive and the size of the country both geographically and demographically does not lend itself to being tackled in the ways described below as in the Hong Kong case study—assessing classroom language for example. This is because of the sheer immensity of the task. It would be physically impossible to train up and standardise classroom language assessors for the whole of the English language teaching cadre in China.

Hong Kong

In Hong Kong, as long ago as the early 1990s, Tam (1992) drew attention to the need for quality control mechanisms for appraising the teaching labour force and proposed an evaluative mechanism. Within the context of secondary school teachers of English, Falvey (1995) discussed the overall lack of training that most secondary school teachers of English had received (of the cohort of 3700 in 1993, only 14.2% were both subject and professionally trained (see also Tsui, 1993). Falvey observed that:

> …. with such a large proportion of the workforce unqualified to teach English, either by subject training or by professional training or both, teachers will, on the whole, have problems implementing any curriculum that requires, in addition to all other general educational knowledge and sound methodological practices, a requisite amount of subject knowledge and pedagogic content knowledge. (1995, p. 2)

Hamp-Lyons and Lumley (1998) described a project in which the development of instruments to measure the writing and speaking proficiency (for English) of all students graduating from the Hong Kong Polytechnic University was under way. The project identified domains to which the criteria of the assessment could be matched, identified actual levels of proficiency of final year students within these criteria and devised descriptions of the identified proficiencies to indicate to end-users what would be *normally* expected of graduates from various academic specialties. These researchers, were, in effect, determining benchmarks of language proficiency for English. The resulting Graduating Students' Language Proficiency Assessment (GSLPA) was first implemented in the 1999–2000 academic year. The test was a wholly workplace-oriented, task-based performance test, designed to specifically assess speaking and writing. Results on the GSLPA are reported on a scale from 1 (low) to 6 (high), with the symbol '+' indicating intermediate points on the scales, such as 2+, 3+. There are altogether 11 bands for the GSLPA (Qian, 2008). In 2007, the Hong Kong Polytechnic University, in collaboration with the then Hong Kong Institute of Education and Lingnan University, developed the Diagnostic English Language Tracking Assessment (DELTA) test, which identifies students' level of language proficiency as well as offering test takers a feedback report (Urmston, Raquel, & Tsang, 2013).

The English language benchmark case study described in this book illustrates how in Hong Kong, the initiative by the Education Commission (through its Report Number 6 in 1995) and the Advisory Committee on Teacher Education and Qualifications (ACTEQ) to set minimum language standards for *teachers* was timely, as other places in both Hong Kong and around the world began to undertake similar initiatives, using increasingly similar assessment methods.

As reported earlier, it was eventually confirmed that a battery of 'formal' tests would be created (i.e. the four skills of reading, writing, listening and speaking), together with a live classroom performance test of classroom language (Coniam & Falvey, 2002). After the tests had been assembled, a pilot study, the Pilot Benchmark Assessment (English) [PBAE], was then held in 1999 and administered to lower

secondary (Years 7–9) English language teachers to see how these coped with the prototype benchmark levels of language ability. The Government of the Hong Kong Special Administrative Region (2000) published the examination syllabus and specifications for the LPATE test in 2000. The first live administration of the tests was held in March 2001.

In early 2000, it was agreed that after 2006 the LPATE would be revised and only offered to new teachers. In addition, and with some attendant controversy, based on the studies that had been carried out, full exemption was offered to teachers holding both a relevant degree and a professional teaching qualification. At the same time, adequate provision (approximately US$30 million) was provided by the HKSAR Government for in-service development courses for teachers to attain the required standards. A revised LPATE was completed in 2007. This was broadly similar to the 2000 version, with some minor amendments (see Urmston Chap. 14 in this volume).

From the beginning, the Hong Kong Examinations and Assessment Authority (HKEAA) administered the LPATE, with the exception of the Classroom Language Assessment component, which is administered by the Government's Education Bureau.

Making the Test Fit the Situation

While some evidence has been given above of countries which have some form of language ability benchmarking in place, the manner in which the assessment of teachers' language ability for English language teaching is conducted also needs to be taken into account. In this context, the test which appears to match closest the demands placed upon English language teachers was the Guam Educators' Test of English Proficiency (Stansfield, Karl, & Kenyon, 1990). The Guam test assessed the four skills of listening, speaking, reading and writing, with some of the test content based around the subject area of language teaching.

While the approach taken by Stansfield et al. (1990) related tasks to the teaching situation, there was still no examination of a teacher's language ability in the language classroom, which may reveal weaknesses not in evidence in a formal test situation. Furthermore, the work conducted by Stansfield and his colleagues occurred before the impetus of performance-based, criterion-related assessment was fully accepted with the result that the majority of the Guam tests contained limited-response items, with a considerable amount of multiple-choice.

Summary

This chapter has outlined a historical picture of research into language benchmarks which were beginning to be investigated and developed in various jurisdictions around the world in the 1990s and 2000s.

Note

1. Penny McKay (1948–2009), the Australian educator, who did much valuable work in this area died at a relatively young age, in the midst of her career. She and her work were honoured and commemorated in a number of articles: http://www.tesol.org.au/files/files/362_Penny_McKay_Article.pdf (Dooley & Moore, 2009).

References

Alderson, J. C., Clapham, C., & Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research, 1*(2), 93–121.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Purpura, J. E. (2007). Language assessments: Gate-keepers or door-openers? In B. Spolsky, & F. M. Hult (Eds.), *The handbook of educational linguistics*. http://www.blackwellreference.com/subscriber/uid=267/tocnode.html?id=g9781405154109_chunk_g978140515410933. Accessed November 2017.

Baird, J. A., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice, 11*(3), 331–348.

Biggs, J. (1996). The assessment scene in Hong Kong. In J. Biggs (Ed.), *Testing: To educate or to select*. Hong Kong: Hong Kong Educational Publishing Co.

Biggs, J. (2012). What the student does: Teaching for enhanced learning. *Higher Education Research & Development, 31*(1), 39–55.

Brindley, G. (Ed.). (1995). *Language assessment in action*. Sydney: NCELTR.

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: A review of the issues. *Language Testing, 15*(1), 45–85.

Burke, B. M. (2013). Looking into a crystal ball: Is requiring high-stakes language proficiency tests really going to improve world language education? *Modern Language Journal, 97*(2), 531–534.

Butler, Y. G., & Iino, M. (2005). Current Japanese reforms in English language education: The 2003 "action plan". *Language Policy*, *4*(1), 25–45.

Carlson, R. E. (1990). Assessing teachers' pedagogical content knowledge: Item development issues. *Journal of Personnel Evaluation in Education, 4*(2), 157–163.

Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System, 28*(4), 523–539.

Chambless, K. S. (2012). Teachers' oral proficiency in the target language: Research on its role in language teaching and learning. *Foreign Language Annals, 48*(4), 604–617.

Chapelle, C. A. (2012). *Reliability in language assessment. The encyclopaedia of applied linguistics*. Chichester: Blackwell Publishing Ltd.

Coniam, D., & Falvey, P. (2002). Selecting models and setting standards for teachers of English in Hong Kong. *Journal of Asian Pacific Communication, 12*(1), 13–37.

Donato, R. (2009). Teacher education in the age of standards of professional practice. *Modern Language Journal, 93*(2), 267–270.

Dooey, P. (2008). Language testing and technology: Problems of transition to a new era. *ReCALL, 20*(1), 21–34.

Dooley, K., & Moore, L. (2009). Penny McKay 1948–2009: A leader in English language education. *TESOL in Context, 19*(2), 50–66.

Eiken Tests. (2017). http://stepeiken.org/overview-eiken-tests. Accessed September 2017.

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing, 10*(3), 235–254.
Elder, C. (1994). Performance testing as a benchmark for LOTE teacher education. *Melbourne Papers in Language Testing, 3*(1), 1–25.
Elder, C., & Kim, S. (2014). Assessing teachers' language proficiency. In A. J. Kunnan (Ed.), *The companion to language assessment*. Wiley-Blackwell: Malden, MA.
Eraut, M. (1994). *Developing professional knowledge and competence*. London: The Falmer Press.
Fahmi, M., Maulana, A., & Yusuf, A. (2011). Teacher certification in Indonesia: A confusion of means and ends. Working paper in economics and development studies. Padjadjaran University, Bandung: Center for Economics and Development Studies (CEDS).
Falvey, P. (1995). The education of teachers of English in Hong Kong: A case for special treatment. In *Teacher Education in the Asian Region. Proceedings of ITEC '95* (pp. 107–113). Hong Kong: Department of Curriculum Studies, The University of Hong Kong.
Farnsworth, T. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly, 10*(3), 274–291.
Fischer, G. (2013). Professional expectations and shattered dreams: A proficiency dilemma. *Modern Language Journal, 97*(2), 545–548.
Fleming, D. (2015). Citizenship and race in second-language education. *Journal of Multilingual and Multicultural Development, 36*(1), 42–52.
Gipps, C. (1994). *Beyond testing*. London: The Falmer Press.
Grant, L. (1995). *Testing bilingual teachers' language proficiency: The case of Arizona*. Princeton, N.J.: Educational Testing Service.
Hajer, A., & Kaskens, A. M. (2012). *Canadian language benchmarks 2012*. Ottawa: Centre for Canadian Language Benchmarks.
Hamp-Lyons, L., & Lumley, T. (1998). *Expectations of exit language proficiency of University Graduates in Hong Kong*. Paper presented as part of 'language assessment in education in Hong Kong' colloquium at the 20th Annual Language Testing Research Colloquium, Monterey, CA.
Han, B., & Qu, X. (2016). A comparative study of five ELT certificates. *Foreign Language Education, 37*(6), 42–47.
Haque, E. V. E., & Cray, E. (2007). Constraining teachers: Adult ESL settlement language training policy and implementation. *TESOL Quarterly, 41*(3), 634–642.
Hu, G. (2004). Building a strong contingent of secondary English-as-a-foreign-language teachers in China: Problems and polices. *International Journal of Educational Reform, 14*(4), 454–486.
Hu, G. (2005). English language education in China: Policies, progress, and problems. *Language Policy, 4*(1), 5–24.
Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics, 25,* 205–227.
Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.
Kimura, Y., Nakata, Y., Ikeno, O., & Andrews, S. (2017). Developing classroom language assessment benchmarks for Japanese teachers of English as a foreign language. *Language Testing in Asia*, *7*(3), 1–14. https://doi.org/10.1186/s40468-017-0035-2.
Kornblum, H., & Garschick, E. (1992). *Directory of professional preparation programs in TESOL in the United States, 1992–1994*.
Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*(2), 131–166.
Li, L., Mazer, J. P., & Ju, R. (2011). Resolving international teaching assistant language inadequacy through dialogue: Challenges and opportunities for clarity and credibility. *Communication Education, 60*(4), 461–478.
Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54–71.
Milanovic, M. (2016). Personal communication October 2016.

McDowell, C. (1995). Assessing the language proficiency of overseas-qualified teachers: The English Language Skills Assessment (ELSA). In G. Brindley (Ed.), *Language assessment in action* (pp. 11–29). Sydney: NCELTR.

McKay, P. (1995). Developing ESL proficiency descriptions of the school context. In G. Brindley (Ed.), *Language assessment in action* (pp. 31–63). Sydney: National Centre for English Language Teaching and Research.

McKay, P. (2000). On ESL profiles for school-age learners. *Language Testing, 17*(2), 185–214.

McKay, P., & Ferguson, R. (2000). English language standards for schools in Australia and China. *Hong Kong Journal of Applied Linguistics, 5*(1), 108–127.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5–12.

Pawlikowska-Smith, G. (2002). *Canadian language benchmarks 2000: Theoretical framework*. Centre for Canadian Language Benchmarks. Retrieved 8 August 2016 from http://eric.ed.gov/?id=ED468319.

Pearson, L., Fonseca-Greber, B., & Foell, K. (2006). Advanced proficiency for foreign language teacher candidates: What can we do to help them achieve this goal? *Foreign Language Annals, 39*(3), 507–519.

Qian, D. (2008). English language assessment in Hong Kong: A survey of practices, developments and issues. *Language Testing*, *25*(1), 85–110.

Reeves, J. (2010). Looking again at add-on ESOL certification. *TESOL Quarterly, 44*(2), 354–364.

Sadtono, E. (1995, April 12–15). *The standardization of teacher trainees in EFL countries*. Paper presented at the 2nd International Conference on Language in Development: The Stakeholders' Perspectives, Denpasar Bali.

Salvatori, M. (2009). A Canadian perspective on language teacher education: Challenges and opportunities. *The Modern Language Journal, 93*(2), 287–291.

Samson, J. F., & Collins, B. A. (2012). *Preparing all teachers to meet the needs of English language learners: Applying research to policy and practice for teacher effectiveness*. Washington, D.C.: Center for American Progress.

Sercu, Lies. (2004). Assessing intercultural competence: A framework for systematic test development in foreign language education and beyond. *Intercultural Education, 15,* 73–89.

Spratt, M. (2015). TKT: Testing knowledge about teaching. In R. Wilson & M. Poulter (Eds.), *Assessing language teachers' professional skills and knowledge* (pp. 242–256). Cambridge: Cambridge University Press.

Stansfield, C. W., Karl, J. & Kenyon, D. M. (1990). *The Guam educators' test of English Proficiency (GETEP).* Final Project Report, Revised. Washington, D.C.: Center for Applied Linguistics.

Sykes, G., & Wilson, S. M. (1988). *Professional standards for teaching: The assessment of teacher knowledge and skills*. Washington, DC: Office of Educational Research and Improvement.

Tam, T. K. (1992). Quality control mechanisms for appraising the teaching labour force. *Education Journal, 20*(1), 17–24.

Tang, C., & Biggs, J. (1996). How Hong Kong students cope with assessment. In D. A. Watkins & J. B. Biggs (Eds.), *The Chinese learner: Cultural, psychological and contextual influences*. CERC and ACER: Hong Kong and Victoria, Australia.

Taylor, C. A., & Angelis, P. (2008). The evolution of the TOFEL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 27–54). New York, NY: Routledge.

Thomas, C. F., & Monoson, P. K. (1993). Oral English language proficiency of ITAs: Policy, implementation, and contributing factors. *Innovative Higher Education, 17*(3), 195–209.

Tsui, A. B. M. (1993). *Report to the Hong Kong language campaign*. Hong Kong: Hong Kong Language Campaign.

Urmston, A., Raquel, M., & Tsang, C. (2013). Diagnostic testing of Hong Kong tertiary students' English language proficiency: The development and validation of DELTA. *Hong Kong Journal of Applied Linguistics, 14*(2), 60–82.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York, NY: Palgrave Macmillan.

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction in the Faculty of Education and Human Development at The Education University of Hong Kong, where he is a teacher educator, working with teachers in Hong Kong primary and secondary schools. His main publication and research interests are in language assessment, language teaching methodology and computer assisted language learning.

Peter Falvey is a teacher educator, formerly a Head of Department in the Faculty of Education, The University of Hong Kong. His main publication and research interests are in language assessment, first and second language writing methodology, and text linguistics.