Journal of Applied Research in Education, vol 19, 2015, p 1 - 14

Half a Century of English Language Assessment in Hong Kong

David Coniam*
The Hong Kong Institute of Education, Hong Kong

Abstract

This paper presents a personal picture of my long-standing association with the English language teaching and assessment situation in Hong Kong. The paper presents a 50-year retrospective of assessment in Hong Kong, through the lens of my own 35-year stint in the territory and my personal experience of English language teaching, teacher education, and assessment. I present a historical and theoretical picture of how English language examinations have moved forward in Hong Kong, and how I was fortunate enough to be involved in the big changes which were taking place in English language assessment in particular. While the picture I portray through this paper is a rather personal one, it contributes to an understanding of how assessment reform has been forward-looking, and largely successful, in Hong Kong, painting a picture of how assessment development has complemented curriculum development. I suggest that it may be instructive for educators in other jurisdictions to consider the long-term picture of development in English language assessment reform in their own country with a view to analyzing their own perspectives concerning the relative success of policy changes and large-scale reforms.

***Keywords*:** assessment, English language, Hong Kong, retrospective

* Corresponding author:
E-mail: coniam@ied.edu.hk

ISSN 1996 – 9945 and ISSN2310 – 3477
Sultan Hassanal Bolkiah Institute of Education
Universiti Brunei Darussalam

INTRODUCTION

This paper details not only my 35-year sojourn in Hong Kong and my long-standing association with its English language assessment situation but also a 50-year journey through assessment in Hong Kong, augmented by my personal experience of years of English language teaching, teacher education, and assessment. I present a historical and theoretical portrait of how English language examinations have moved in Hong Kong – onwards and upwards, with me fortunate to be in the middle when big changes were taking place in English language assessment: from how assessment was conceptualised to how it was delivered.

The picture I illustrate is a personal perspective, based on my experiences and perceptions of the Hong Kong situation. The major issues discussed, however, will reflect development in English language teaching and assessment that many jurisdictions in Asia have been grappling with over the past half century in terms of different types of curriculum and assessment reform and the extent to which such reforms have been embraced. Pictures of curriculum reform have been published for a number of Asian countries. Ho (2002) presented snapshots of different countries in East Asia, while specific country analyses were reported by Boyle (2004) on Hong Kong, Wenfeng and Lam (2009) on China, and Choi (2015) on Japan and South Korea. The current paper adds to our repository of knowledge, complementing the picture of curriculum development, by providing a specific blueprint of assessment development in one jurisdiction.

To provide an anchoring backdrop of my experience, at first I frame issues within the context of the key test quality concepts *Validity*, *Reliability*, and *Washback*. The paper then moves through assessment in Hong Kong one decade at a time, with my experiences framed as appropriate against a relevant key test quality concept. The paper closes by making reference to how such reflection may be conducted profitably in a broader Asian context, with a view to gauging development in different countries and regions.

The backbone of English language assessment in Hong Kong is the public examinations body, the Hong Kong Examinations (and Assessment) Authority (HKEAA). The HKEAA was established in 1977, prior to which, public exams had been administered under the aegis of the then Education Department(ED) (see Choi & Lee, 2010, p.60). I have had a long-standing association with the HKEAA, and in fact still do.
My formative years were spent there in the late 1980s. Consequently, a considerable amount of my presentation in this article relates to my experience and association with the HKEAA.

TEST QUALITY CONCEPTS

One of the major issues that first needs consideration is: What is the purpose of the English language curriculum? 40 years ago, the 'purpose' of an English language curriculum might have been framed as:

1. Mastering every grammatical structure, and
2. Testing what students knew about English grammar.

Through major curriculum development from the late 1970s onward with the advent of a Communicative Approach to ELT (e.g., Littlewood, 1981), the focus shifted from a single focus on structure to one involving a more cognitive, affective humanistic approach, with 'communication' being a goal as important as structure (see for example, Richards and Rodgers, 2001).

The current 'purpose' of an English language curriculum may then be framed as:

1. Students communicating in English, and
2. Testing what students can *do* in English.

The key concept here is *Validity*. Validity (see, for example, Bachman & Palmer, 1996; Messick, 1989) may be framed as:

1. What 'skills', 'abilities', and 'constructs' are being tapped in the test, and
2. How far a given test score can be interpreted as an indicator of the abilities or constructs to be measured?

The second key concept is *Reliability* (see, for example, Hughes, 2003), which relates to how results awarded to test-takers change across periods of time, across different groups of students, and between markers, viz.:

1. The degree of objectivity in a test, with subjective tests generally having lower reliability than more objective tests,

2. Test length (i.e., how many items there are in the test), with longer tests generally being more reliable, and
3. The amount of question choice test that takers have.

The third key concept is *Washback* (see, for example, Alderson, 2004; Cheng, 2005; Choi & Lee, 2010), which relates to the effect that changes to examinations have on teaching. Despite the fact that the HKEAA has been an independent examinations body, positive washback has been at the forefront of many of the major changes that have occurred with its English language examinations, and it has taken very seriously the notion that examinations should encourage worthwhile classroom practices.

In order to make the material more digestible, and to put issues into perspective, I will frame issues as I move through the paper through the lens of decades. Although I was not in Hong Kong in the 1960s and 1970s, here I have managed to gain access to past ED/HKEAA documentation to fill in the gaps.

THE 1960S AND 70S – BEHAVIOURIST TIMES

The 1960s and 70s were Behaviourist times. In line with Behaviourist principles there was a strong focus on reliability, and accuracy was the order of the day (see Howatt, 2004). In tandem with the 'methodology' underpinning Behaviourist principles, the activities that predominated were translation, grammar drills, and a substantial amount of multiple choice questions. I will illustrate with a couple of examples.

One section of the 1967 School Certificate Examination, English Paper III required candidates to translate from English into Chinese. Figure 1 shows a sample.

Figure 1. Passage for translation (extracted from the 1967 School Certificate Examination)

> SECTION B (30 marks) Translate the following passage into Chinese
>
> *On a splendid September day, I left my native land. After a very interesting journey I arrived in London. I was amazed at the difference between my small village and the huge city. What traffic! What an uproar in the streets! At first the noise nearly deafened me, but after an hour or so I became used to it. Everything was new and strange and I must confess that my first impressions were not very favourable. When I arrived at the hotel which had been chosen for me by a friend, I felt very tired. I had seen a great deal in one day, and I felt in need of rest in mind and body that evening.*

While the text is clearly dated, and there are (from a current perspective) some non-politically correct elements ('my native land'), there are a number of points worth-considering.

1. It is not a young person's text. It is written by a middle-aged examiner ('I must confess that my first impressions'; 'I felt in need of rest in mind and body'). This is not how a young person speaks – or spoke 50 years ago. Times have changed and the genre and makeup of texts presented to 16-year-olds are now more relevant to them (see, for example, Krashen & Terrell, 1983 on the 'Natural Approach').
2. Testing points have been chosen to assess a range of elements: past tense, passive, 'What a …', relative clauses
3. It is not a spoken text, although it tries to appear that way. There are complex sentences and embedded relative clauses.

Another feature of past English language examinations was a clear focus on grammar. The sample in Figure 2 below – from the 1966 Secondary School Entrance Examination (SSEE) – required candidates to transpose sentences from indirect speech into direct speech. This is a fascinating exercise in that it is one that would almost never take place in regular communication, either written or spoken. Validity in this task is, consequently, very low.

Figure 2. Indirect speech into direct speech exercise (extracted from the 1966 SSEE)

a. The old man said that he had seen the doctor.
a. “ _____”, the old man said.
b. The teacher said that we were being very naughty.
b. “ _____”, said the teacher.
c. She asked me when he would arrive.
c. “ _____”, she asked me.
d. He told John to open the door.
d. “ _____”, he said to John.
e. We asked the girl what her name was.
e. “ _____”, we asked the girl.

In line with a focus on reliability, there was a strong emphasis on multiple-choice testing – that first appeared in Hong Kong English language examinations in 1969 (King, 1994).

In order to eliminate the possibility of candidates cheating, multiple forms of the same test were created, with different lettering for the options (ABCD, EFGH, WXYZ), reordered options, and the key placed in a different place. Figure 3 delineates a mockup sample.

Figure 3. Mockup of multiple-form MC test

1. The _____ of sleepless nights made her ill.
A. strain [KEY]
B. stress
C. tension
D. pressure

4. The _____ of sleepless nights made her ill.
Q. stress
R. strain [KEY]
S. tension
T. pressure

9. The _____ of sleepless nights made her ill.
W. strain [KEY]
X. stress
Y. tension
Z. pressure

A major issue with assessment in the 1960s and 70s was the effect on teaching caused by the format of the examinations (MC in particular) and negative washback (see Alderson & Hamp-Lyons, 1996). While MC testing was an accepted part of the culture, the HKEAA was concerned about its negative washback, and strove to minimise the fallout by trying to restrict the amount of MC practice paper work that schools might do for the English language public examinations. They performed this by not publishing the MC papers from the examinations (King, 1994). Unfortunately, this did not prevent teachers from getting hold of the papers since a classic workaround was for a teacher to ask each student in their class to memorize two or three MC questions and to write them down for the teacher immediately after the examination.

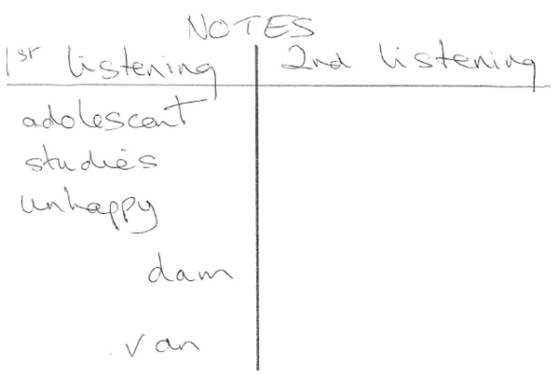
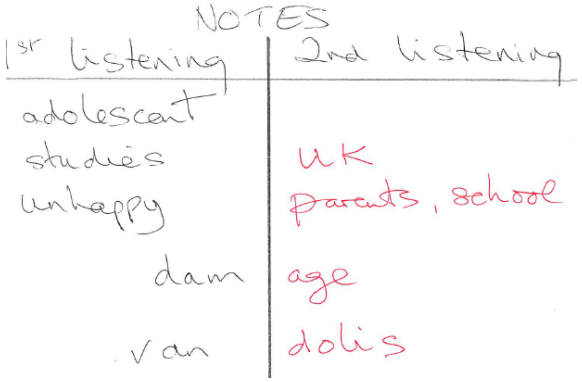
In the 1970s, while there was a continued focus on reliability, meaning and relevance were beginning to enter English language examinations. The Year 13 Hong Kong Advanced Level Examination (HKALE) Use of English (UE) exam was very reliability-focused, though it was 'relevant' to tertiary-level studies. Elements in the exam included cursory reading (albeit multiple-choice) and an academic lecture listening test.

Use of English Listening and Oral Tests: Format and Teaching Approach

It is worth dwelling awhile on the format and teaching approach that characterized the first series of UE listening tests. The format in which the test was delivered was somewhat less natural than the current format where candidates have time to look over the question booklet before the listening input, having the opportunity to focus on what the test will be about before they hear (usually once) the tapescript. In those days, the listening test was played twice; the candidates listened 'blind' in that they did not receive the question booklet until after having listened twice. The intended objective was that – as with an academic lecture – they would take notes, and later make sense of them to answer the questions.

Unfortunately, this objective (and test validity along with it) was, however, widely circumvented by the 'two-pen approach' devised by smart (test-wise) Hong Kong teachers. Under this 'two-pen approach' (see Figure 4 below), candidates took a blank sheet of paper, which they divided in half vertically with a line down the middle. During the first listening, they made notes down the left-hand side in blue. On the second listening, they took a different coloured pen (e.g., red) and made notes down the right-hand side of the page in red. Finally, they opened the question booklet, and attempted to patch their notes together to answer the questions (see King, 1994). The activity was more like speed dictation than a listening test where meaning was being assessed in the context of a stream of speech. Consequently, the listening test in this format was low in validity and was one reason for the subsequent overhaul of the UE examination.

Figure 4. Mockup of use of two-pen English Listening test

1st listening: Make notes on the LH side (in blue)	2nd listening: Make notes on the RH side (in red)
	

The format of the Hong Kong Certificate of Education Examination (HKCEE) English oral test that ran from 1974 to 1995 was reliability focused, with emphasis on accuracy rather than fluency. This was evident in the first part *Reading a dialogue*, where the candidate and examiner read a dialogue together. The more open Part 3 – *Conversation with the examiner(s)* – in principle allowed for fluency work, but in practice this was more akin to an 'inquisition' in the manner in which two examiners tended to 'interrogate' a candidate.

THE 1980S – A COMMUNICATIVE APPROACH TO LANGUAGE TEACHING

In the context of a worldwide movement that advocated that there was more to language teaching than merely grammar (Hymes, 1972), the 1980s saw the advent of a Communicative Approach to Language Teaching.

In line with the principles of a 'Communicative Approach' – and the needs of society/business expanding – elements more than grammar began to come into focus in both school curriculums and English language examinations. There began to be a greater focus on language use, which in examination terms meant greater validity, in that the examination score gave more of an indication of what candidates could ***do*** in English than previous examination formats did (Messick, 1989).

The effect of the new communicative movement was major revisions to the key HKCEE and HKALE (see Choi & Lee, 2010). Multiple-choice and grammar testing were still part of the public examinations, but were being quietly de-emphasized.

One major innovation to the HKCEE of English language was the introduction, in 1986, of a listening test. The HKCEE Listening Test was more general in its orientation than the UE Listening test, which, being HKU's entry test, had the format of an academic lecture. This was a major commitment by the HKEAA. Since the radio signal was not strong enough to cover the whole of Hong Kong, the listening test involved the school halls of most secondary schools being equipped with 'induction loops'. Nevertheless, the listening test still required five parallel sessions, and the HKEAA had 25,000 sets of headphones for each session (see King, 1994).

In 1989, the UE exam was completely revised (see King, 1994). It was no longer solely HKU's entrance test, as it had been, but was intended to perform the dual role of a tertiary entrance test and be a valid assessment for Year 13 school leavers who would join in the workplace, working for a company or business.

The 1989 revision of the UE examination was therefore much more communicative in its orientation – with better validity; there was still a reliability focus, however. Nonetheless a major focus of the HKEAA's was the effect of washback: having students do things in their English language classrooms that would have wider relevance than merely a university entrance test. One major omission in the washback picture, though, was the fact that there was still no oral test in the UE – a situation that was not rectified until 1994.

At this point, thus, we can make a major statement. It is that the effect of the examination on teaching *did* matter in the 1980s. It was a major concern of the HKEAA's. It would be interesting to compare this 'imperative' in the context of other Asian nations or jurisdictions in the 1980s, some of whom are still trying to incorporate some elements of a communicative approach to teaching and testing in their school English language curricula and examinations even now, in the 2010s (see Choi, 2015, for a comparative discussion of the cases of Japan and South Korea).

THE 1990S – FOCUS ON LANGUAGE USE EXTENDED, FOCUS ON STANDARDS

Focus on Language Use

Following greater adoption of the principles of a 'Communicative Approach', language use began to come more to the fore. As a consequence, the focus on grammar, as well as on multiple-choice, became to be somewhat less emphasized.

In 1996, the HKCEE of English language underwent radical revision. The previous listening test, which had all been multiple-choice, was incorporated into an 'integrated' listening/reading/ writing paper. Cheng (2005) investigated this change from the perspective of washback and reported how the modification of an examination changes teachers' classroom practice. She stated that changes in teaching content were the most obvious indicators, with changes to classroom activities in line with the new more communicative examination being observed (Cheng, 1997, p.49).

The oral component of the examination was radically revised as well. It changed from a mainly 'interrogation' format to one where the major part was a group discussion. Previously the examiners had dominated the discussion; in the revised version of the oral the examiners were only assessors, taking no part in the discussion at all. Validity was therefore enhanced but to maintain reliability; considerably more training and standardization was required and provided.

Focus on Standards

The mid-1990s also saw a strong focus on standards – teacher English language standards in particular.

A major initiative by the Government of the Hong Kong Special Administrative Region (HKSAR) involved establishing minimum language proficiency examinations (also known as 'language benchmarks') for all teachers in Hong Kong primary and secondary schools. The genesis of these benchmark examinations lay in concern – expressed since the early 1990s by different sectors of the business and education communities in Hong Kong – over perceived falling language standards especially after the publication of research conducted in the early 1990s that revealed that less than 20% of the secondary workforce of Hong Kong's English language teachers were both academically and professionally qualified (Tsui et al., 1994). The government therefore deemed it essential that teachers of English developed their second language skills as one of the prerequisites for being able to teach and adapt to new assessment methods and curricular objectives in their classrooms.

Of the full cohort of 3,700 secondary school teachers of English in 1993, only 14.2% were both subject and professionally trained. Many teachers of English in secondary schools had received neither subject content nor professional training or were teachers of other subjects, forced to teach English merely because of a shortage of qualified staff. One major initiative by the Education Commission (established 1982) requested the Advisory Committee on Teacher Education and Qualifications (ACTEQ) to investigate the specifying of minimum language proficiency standards (the 'Language Benchmark' test as it was initially known, and subsequently the 'Language Proficiency Assessment of Teachers of English' [LPATE]) (Education Commission, 1996, p. 11).

The 1996 Consultancy Study and Follow-up

In early 1996, a study in which I was heavily involved as one of the two co-principal investigators was commissioned by the Education and Manpower Bureau (EMB) to investigate the feasibility of benchmarks for teachers of English language. The consultancy team reflected a broad spectrum of expertise and experience from local and international language teachers and language teacher educators at primary and secondary levels, including as many different stakeholders as possible. Survey data was collected at both local and international levels, with responses indicating widespread agreement for the establishing of minimum-standard language assessment (see Coniam & Falvey, 1999a).

An initial test battery that was constructed to assess teacher English language standards comprised a three-part paper-and-pencil formal test component, an oral component, and an observation of two live lessons (the Classroom Language Assessment performance test of an English teacher teaching two English lessons). The latter test was considered to be the most valid part of the test battery since it consisted of a performance test during a genuine *target language use* situation (see Coniam & Falvey, 1999b).

The English language benchmark subject committee: Purpose and brief

Following the 1996 Consultancy Study, the next phase in developing English language proficiency standards was undertaken by a broad-based committee, representative of all stakeholders in the teacher, teacher education, and education fields in Hong Kong. This committee, the English Language Benchmark Subject Committee (ELBSC), was convened in October 1997 under the auspices of the Hong Kong Examinations (and Assessment) Authority (HKEAA) to produce language proficiency standards specifications and an assessment syllabus for promulgation to Hong Kong teachers of English language prior to a large-scale pilot exercise – the Pilot Benchmark Assessment (English), known as the *PBAE*.

Classroom language assessment

One major objective in developing the Hong Kong Classroom Language Assessment (CLA) criterion-referenced scales (with accompanying descriptors) was the desire for transparency so that teachers and informed lay-persons, with appropriate training, could reach similar grades when viewing videos of English teachers and rating them on the four CLA scales.

The Pilot benchmark assessment (English)

The PBAE ran from November 1998 to January 1999, and involved large-scale testing of all the assessment instruments proposed and developed by the ELBSC – including two assessments (as per the requirements of the CLA component of the test) of teachers teaching their own English language classes.

Prior to the PBAE, the ELBSC felt that no exemptions from any of the LPATE tests should be allowed. The results of the PBAE were, however, surprising and, to an extent, gratifying in relation to results, qualifications, and relevant background. As a result, those who had both a relevant background and qualifications were exempted from having to sit the LPATE (see Coniam & Falvey, 2003).

Going live: The first administration of the LPATE

The LPATE syllabus was published in mid-2000, and a series of six seminars attended by approximately 10,000 teachers were held to explain the government policy. This was, however, the first time that the Professional Teachers' Union had been able to comment on the LPATE issue publicly. Consequently, the seminars managed to convey little of the spirit of the government's intention to upgrade the English of the teaching profession. Rather, the LPATE was viewed by teachers, especially primary school teachers, as a stick with which Government intended to beat English language teachers. In the years since its first administration in 2001, the LPATE has been investigated and discussed by a number of researchers – from the perspectives of the test's advantages (McGrath, 2000), as well as its perceived problems (Glenwright, 2002; Glenwright, 2005). Whether or not standards as currently set actually reflect the need of the current English language teaching profession in Hong Kong is a major issue that needs further investigation.

The 2000s – SARS, '3+3', ONSCREEN MARKING

Returning to the chronological timeline, three issues dominated the 2000s. The two that stood out were major changes to both the Hong Kong education and examination systems. A third issue was the government and public reaction to Severe Acute Respiratory Syndrome (SARS).

Before 2009 – when the education system underwent major curriculum and examination reform – Hong Kong's education system was modelled on the British system. Secondary schools operated on a 5+2 model with students being streamed ('banded') into three broad bands of ability, each band covering approximately 33% of the student ability range. Public examinations in Hong Kong were conducted by the Hong Kong Examination (and Assessment) Authority (HKEAA). Prior to 2012, there were two major public examinations. The Hong Kong Certificate of Education Examination (HKCEE) was administered at the end

of eleven years of education – Secondary 5 (Year 11). The total candidature for the HKCEE was in the region of 100,000, of whom approximately 80,000 were school candidates. At the end of Secondary 5 (Year 11) students could continue in full time education for two more years – although there were only places for approximately 38% of the Year 11 cohort to continue on to Year 12 and 13 studies. At the end of Secondary 7 (Year 13), students sat the Hong Kong Advanced Level Examination (HKALE), which was also used for university entrance purposes. In 2007, the total candidature for the HKALE was approximately 36,000.

The Hong Kong secondary school curriculum underwent significant restructuring in 2009. Under the restructuring, secondary education now lasts six years with a single public examination (the Hong Kong Diploma in Secondary Education [HKDSE]) administered at the end of Year 12 (age 18).The annual candidature in 2014 was approximately 80,000. The corollary is that many more students now go on to Year 12 than went on to Year 13 previously – before the changes, the annual HKALE candidature was in the region of 40,000.

In line with the public examinations about to undergo drastic structural change in 2009, the examinations themselves (of which English language was at the forefront) – with an eye to the 2009 curriculum restructuring and the new HKDSE in 2012 – also saw massive changes to examination content and format, to marking, and to grading in 2007.Onscreen marking (OSM), which I will discuss below, was another key innovation which began to come in in the 2000s.

The other major event of the 2000s, as mentioned, was Severe Acute Respiratory Syndrome (SARS), which had a considerable impact upon the workings of the Hong Kong education and assessment system, as will now be described. To prevent an outbreak of the disease throughout the education system, classes in Hong Kong were suspended during most of April 2003. As a precautionary measure, all students and teachers were required by the Education and Manpower Bureau (EMB) to wear a face mask after the reopening of schools in late April 2003. While the use of face masks posed some level of discomfort to wearers, the effect as far as classroom settings were concerned was that teachers and students had to interact with some facial cues removed. From an assessment perspective, the situation was then exacerbated in the public examination, the Grade 11 Hong Kong Certificate of Education Examination (HKCEE) oral test held in the month of June, because everyone involved – examiners, test-takers, and administrative staff – had to wear face masks at all times.

There was great concern that the wearing of a facemask would invalidate some of the assessment results. To investigate whether wearing a facemask intruded on the oral assessment score, I conducted a study in March 2004. In the study, the entire Secondary 5 cohort of a Band 2 (average ability) Hong Kong secondary school took a past HKCEE oral test both with and without face masks – five classes, a total of 186 students who were sitting the test as their mock HKCEE oral examination. Conditions of the public examination were replicated as far as possible: grouping unfamiliar students together, avoiding reuse of test materials, and adhering to standard examination practice concerning room conditions, time for preparation, and examination timing. The number of raters, their training, and their behavior during the actual test also followed standard HKEAA procedures.

Contrary to expectations and much to the delight of the authorities, test data results did ***not*** suggest that face masks had an effect on test-takers' oral test scores. Severity/leniency differences were apparent in all the facets modelled using multi-faceted Rasch analysis – raters, bandscales, and prompt materials – apart from the face mask condition, on which facet no difference emerged. Similarly, non-significant results emerged on t-test analyses conducted for all bandscales used – even the *Audibility* and *Comprehensibility* bandscales which would have been most susceptible to the effect of the face mask. Whereas the wearing of facemasks appeared to have a deleterious effect on validity, reliability was – surprisingly to many – not affected.

2007 New HKCEE English Language Exam

The revisions of the HKCEE English language examination in 2007 presaged major changes that were to be implemented for all subjects with the advent of the HKDSE in 2012.

This was the biggest 'upheaval' ever for English language examinations, and a number of major changes were implemented. In an interesting adjustment of policy, there was a much greater rapprochement between the HKEAA and the Curriculum Development Institute (CDI) than in the past. Effectively the new syllabus was produced by the HKEAA in conjunction with the CDI.

The English language examination became standards-referenced (as opposed to the strict norm-referencing that had long dominated the Hong Kong examination system), school based assessment was introduced, and the two English language syllabuses, Syllabus A (originally for Chinese medium schools) and Syllabus B (originally for Anglo-Chinese [English-medium] schools) were collapsed into a single examination.

There were also significant changes to the format of the English language examination, whereby on each examination paper, a single theme (schema) ran through the paper, rather than previously a paper consisting of a set of unrelated subtests.

From a number of perspectives, the validity of the English language examination has been enhanced. The assessment for learning side of School Based Assessment (SBA) made it possible for students to relate more to their peers and the material in the examination than it was with having to sit an examination with a bunch of strangers in an examination hall. Given the long history of norm-referencing, and the more 'humanistic' approach of criterion referencing, a question raised was whether standards would 'slip' as more students achieved potentially higher marks. This did not occur, however. HKCEE pass rates remained pretty constant, and markers did not appear to have 'overmarked'.

OSM in HK

As stated earlier, another major change to the Hong Kong assessment horizon was the manner in which examinations were marked, with, initially, the new English and Chinese syllabuses being marked on screen. Since this has been such a major change to how examinations are marked in Hong Kong, and is an area where Hong Kong is effectively leading the world, the following sections outline a number of studies for English that I and others conducted to investigate the validity of OSM in Hong Kong. Both quantitative and qualitative studies were conducted, investigating a range of issues among which have been: statistical comparability, marker reactions to OSM and to the system, and marker technological readiness.

The discussion below mainly addresses the research questions from the perspective of the English language examination in the Year 11 HKCEE.

Statistical Comparability

Background to the 2008 English study

The largest data set for the English studies was drawn from the Writing paper of the 2007 Hong Kong Certificate of Education Examination (HKCEE) English language examination, where the candidature was 99,771. In this examination, candidates had to complete two writing tasks. Task 1 was a guided narrative piece of writing requiring approximately 150 words. Task 2 was an open-ended task requiring approximately 250 words, on which candidates had a choice of two questions: the first descriptive, where candidates had to explain why they would like to work in the fashion industry; the second argumentative, with candidates having to put the case for whether it was more important to be clever than beautiful (HKEAA,2007, p. 18). All scripts were double marked.

Reliability on the Writing paper was monitored through inter-marker correlations as well as correlations with other papers and with the subject mark for the whole HKCEE English language examination (King, 1994, p. 6). For 2006 – when PBM was still the modus operandi – the inter-marker correlation (i.e., 188 markers, with each marking about 800 scripts) was 0.79. The correlation of the Writing paper with the subject mark for the whole examination in 2006 was 0.89.

A high correlation is generally taken to be 0.8 or better (see, for example, Hatch & Lazaraton, 1991, p. 441). Correlations between the HKCEE Writing paper and the other papers were generally high. While the correlation between the Writing paper and the Speaking paper was somewhat lower at 0.72, the correlation with the School Based Assessment paper was high at 0.83. Finally, the correlation with the whole subject mark for 2007, when OSM was adopted, was a high 0.90, very comparable to the 2006 figure. An immediate observation here was that the introduction of OSM did not impact on test reliability.

The English study – data

One hundred ninety six markers marked the 2007 HKCEE Writing paper Task 1B2, of whom 117 (59.7%) were experienced markers and 79 (40.3%) first-time markers. 46 of these 196 were identified as potential markers for the current study on the basis of two criteria: first, that they had good marking statistics in their marking of the 2007 HKCEE Writing paper, for example, inter-marker correlations and high correlations with the objectively-marked Reading paper. Second, as far as possible, the sample would be a representative cross-section of markers in terms of gender and qualifications, as well as their teaching and marking experience. For these first-time markers, the 'new' experience would, conversely, be paper-based marking. Of the 46 potential participants shortlisted, however, only 6 new markers with good statistics were identified. 30 markers were eventually recruited to take part in the study – 5 (16.7%) new and 25 (83.3%) experienced markers. Each marker marked 100 scripts. They were informed they would be marking some scripts from the 2007 HKCEE examination, and that their batch of 100 scripts would contain some of the scripts they had marked previously. They were not informed that they would essentially be re-marking 100 scripts which they had previously marked (see Coniam, 2009 for a description of this procedure).

The total sample therefore comprised 3,000 scripts, of which there were 2,145 different test-takers. Care was also taken to ensure that scripts selected from each marker's batch represented the full range of levels (i.e., 1 to 6) of the subscales. As mentioned, analysis was conducted using both classical test statistics such as inter-marker and inter-paper correlations (King, 1994, p. 6).

The English study – Results and discussion

An analysis of the two prompts (Coniam, 2009) indicated that while there was a significant difference between the mean score for the two prompts, this could be attributed to test-taker ability rather than to the effect of the prompt. Importantly, however, both prompts exhibited very comparable means under the different marking conditions OSM vis-à-vis PBM. T-test results for the two methods of marking were not significant for either prompt, suggesting that the prompt did not contribute bias to the analysis.

Discrepancies between the two forms of marking

A common criterion for invoking re-marking (i.e., the use of a third marker) has been established as two markers differing from each other by more than one score point on a 6-point scale (see, for example, Attali & Burstein, 2005, p.13). Compared with the overall discrepancy rate for the 2007 HKCEE Writing paper of 10%, the study revealed a lower incidence of discrepancies between the two forms of marking, with an overall figure of 8.1%. While the total number of scripts receiving higher scores was slightly higher at 4.6% when marked in OSM as against a slightly lower figure of 3.5% being more severely marked, the figures were quite close with no significant differences reported on t-tests. The incidence of discrepancies +/- 5 points emerged as very similar on both topics.

Marker Technological Competence and Attitude towards OSM

To more fully explore attitudes towards OSM, in Coniam's (2009) study, markers completed a post-marking questionnaire detailing their attitudes towards the onscreen and paper-based marking processes. The questionnaire was in three sections, with the first involving background demographics. The second,

computer familiarity issues, concerned markers' computer proficiency, how competent they were at manipulating the mouse, enlarging and scrolling the screen image, and ergonomic issues such as desktop height and screen resolution. The third, marking issues, tapped issues such as their perceived accuracy for onscreen / on-paper marking, how tired their eyes became through marking in the two modes and how often they needed to take a break while marking. It also enquired about their preference as to marking mode, and whether they preferred marking at home or at a marking centre.

Questions were posed on a 6-point Likert scale, with '6' indicating a positive response or agreement, and '1' a negative response or disagreement. Markers were also asked to provide written comments on any aspect of the OSM process that they wished to comment on.

On the question of how proficient markers were, responses indicated that markers felt themselves to be quite competent technologically. Markers responded to questions about ergonomic issues such as screen height and resolution positively, although new markers were, in general, more positive than experienced ones. Having to travel to a special marking centre was also reported as less of a problem by new markers. A similar finding was recorded in the preference for marking at home or at a centre, where new markers rated centre marking significantly more positively then did experienced markers. Overall, however, despite certain misgivings, it could be seen that even experienced markers were aware of the potential benefits available with OSM – rather than the new system simply inspiring difficulties and drawbacks.

The picture that emerged from the different OSM studies presented above was that buy-in and acceptance by markers clearly increased with each year. In 2012, with all examinations marked on screen, it was very important to ensure that the system was reliable. The studies I steered reveal that this was likely to be the case. It was very rewarding to have been part of this validation process for onscreen marking – in which Hong Kong is a world leader.

IN CLOSING

As will be clear from my account in this paper, Hong Kong English language examinations have come a long way in 50 years – not only theoretically, but technologically and practically. In large part, I developed with the English language examinations that I had contact with: from a mechanistic, behavioral orientation to one which is more humanistic, more thinking, more feeling, one which is in tune with the time, with research, and with both assessment and teaching.

While the picture I have portrayed throughout this paper has been a personal one, it contributes to an understanding of how assessment reform has been quite forward-looking, and largely successful, in one jurisdiction – Hong Kong. The picture of assessment development in the paper has been intended to complement that of curriculum development. As such, it may be instructive therefore for educators in other countries and jurisdictions to consider the long-term picture of development in English language assessment reform in their own country with a view to analyzing where they stand in terms of the success of policy and of being in tune with current thinking.

REFERENCES

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*, 280-297.

Alderson, J.C. (2004). Foreword. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. ix-xii). Mahwah, NJ: Lawrence Erlbaum Associates.

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback – a case-study. *System, 30*(2), 207-223.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford, NY: Oxford University Press.

Boyle, J. (2004). Linguistic imperialism and the history of English language teaching in Hong Kong. In K. K. Tam &T. Weiss (Eds.), *English and globalization: Perspectives from Hong Kong and mainland China* (pp. 65-84). Hong Kong: The Chinese University of Hong Kong Press.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education, 11*(1), 38-54.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge, UK: Cambridge University Press.

Choi, C. C., & Lee, C. (2010). Developments of English language assessment in public examinations in Hong Kong. In L. Cheng &A. Curtis (Eds.), *English language assessment and the Chinese learner*(pp. 60-76). Routledge: New York & London.

Choi, T. H. (in press). Glocalization of English language education: Comparison of three contexts in East Asia. In C. M. Lam & J. Park (Eds.), *Sociological and philosophical perspectives on education in the Asia-Pacific region* (pp. xx-xx.). Hong Kong, China: Springer.

Coniam, D., & Falvey, P. (1999a). Setting standards for teachers of English in Hong Kong – the teachers' perspective. *Curriculum Forum, 8*(2), 1-27.

Coniam, D., & Falvey, P. (1999b). The English language benchmarking initiative: A validation study of the Classroom Language Assessment component. *Asia Pacific Journal of Language in Education, 2*(2), 1-35.

Coniam, D. (2002). IT use in the English language classroom in Hong Kong: How far is Government policy being achieved? *Education Journal, 30*(2), 21-39.

Coniam, D. (2009). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation, 15*(3), 243-263.

Coniam, D. (2013). The increasing acceptance of onscreen marking – the 'tablet computer' effect. *Journal of Educational Technology & Society, 16*(3), 119-129.

Coniam, D., & Falvey, P. (2003). Benchmarking the benchmark: assessing the fit of a new test with its target population of teachers of English in Hong Kong. *Hong Kong Journal of Applied Linguistics, 8*(1), 1-15.

Falvey, P., & Coniam, D. (2010). A qualitative study of the response of raters towards onscreen and paper-based marking. *Melbourne Papers in Language Testing,* 1*5*(1), 1-26.

Ho, W. K. (2002). English Language Teaching in East Asia Today: An Overview. *Asia Pacific Journal of Education, 22*(2), 1-22.

Howatt, A. P. R. (2004). *A history of English teaching* (2nd ed.). Oxford, NY: Oxford University Press.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, UK: Cambridge University Press.

Hymes, D. (1972). On communicative competence. In J. B. Pride (Ed.), *Sociolinguistics* (pp. 269-293). Middlesex, UK: Penguin Books.

King, R. (1994). Historical survey of English language testing in Hong Kong. In J. Boyle & P. Falvey (Eds.), *English language testing in Hong Kong* (pp. 3-29). Hong Kong: The Chinese University Press.

Krashen, S., & Terrell, T. (1983).*The natural approach: language acquisition in the classroom.* Oxford, NY: Pergamon Press.

Littlewood, W. (1981). *Communicative language teaching*: *An introduction*. Cambridge, UK: Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Richards, J. C., & Rodgers, T. S. (1986/2001): *Approaches and methods in language teaching: A description and analysis*. Cambridge, UK: Cambridge University Press.

Tsui, A. B. M., Coniam, D., Sengupta, S., & Wu, K.Y. (1994). Computer-mediated communication and teacher education: The case of TELENEX. In N. Bird, P. Falvey, A. B. M. Tsui & A. McNeill (Eds.), *Language and learning.* Hong Kong: Government Printer.

Wenfeng, W., & Lam, A. S. L. (2009). The English language curriculum for senior secondary school in China: its evolution from 1949. *RELC Journal, 40*(1), 65-82.

BIODATA

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction at The Hong Kong Institute of Education, where he is a teacher educator. His main publication and research interests are in language assessment, language teaching methodology and computer assisted language learning.