Chapter 3
Issues in High-Stakes Assessment

David Coniam and Peter Falvey

Abstract This chapter describes and discusses the major issues involved in high-stakes assessment and refers, where appropriate, to the language benchmark case study, which is described in the following chapters. The full taxonomy of major issues outlined below is not and need not always be present in its entirety in every set of benchmarks, including language benchmarks. However, most major issues need to be taken into account whenever agencies and assessment specialists meet to plan, create, establish and implement benchmarks either for the public or for specialist bodies.

Philosophical Perspectives

Chapter 2 discussed the changing paradigm in testing and assessment of all types (e.g. school, public examination, vocational assessment, etc.). When involvement in a high-stakes assessment procedure consists of stakeholders such as government, government agencies and assessment specialists, it is vital that government agencies be involved and well-briefed from the beginning. One reason for this is that government officials may not be familiar with changing paradigms or current assessment techniques. They have often been educated in an assessment environment far different from that prevailing at the time of a new assessment initiative. It is then necessary to determine how far the government, and its agencies can accept the methods proposed by the assessment specialists within the policy parameters in which they work.

In addition, government and its agencies, working together with specialist assessment consultants, are able to consider policy issues that are far broader and more far-reaching than the narrow focus which the assessment specialists, by themselves,

D. Coniam · P. Falvey (✉)
Department of Curriculum and Instruction, Faculty of Education & 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_3

bring to the task. Often this leads to questions such as ‘How much of what is being proposed can be achieved’? and ‘How much is it all going to cost’?

Even more importantly, after an initial briefing and after agreement on the philosophical stance to be adopted, it is essential that there is ongoing dialogue between the government, its agencies, assessment specialists and other key stakeholders. This is because whatever may be proposed by the assessment specialists, implementation of a high-stakes assessment procedure which involves government and its agencies must fit the government policy of the time. A project which is developed over a number of months or years is often subject both to changes of government personnel (who always require additional orientation and briefing) and government policy. These changes can often frustrate specialist assessment experts. However, because policy supersedes whatever advisors might propose, advisors must learn to accept policy, personnel changes and the sociological context of the project.

Policy Considerations—Washback

Linked to the issue described above are policy decisions affecting the washback effect of high-stakes forms of assessment. The term *washback* is used in assessment to indicate that the creation of test types, test questions or test specifications will produce an effect which will wash back from the test developers to the test takers so that test taker behaviour is affected (for a detailed description of the term *washback* see Cheng & Curtis, 2012). A simple example is the effect on proficiency tests of the introduction of an oral component into a battery of tests that formerly did not contain one. The introduction of the oral test will have an immediate effect on the behaviour both of the test takers and on those who run courses for the test takers.

In the case study, described in Chap. 6, a number of policy decisions were made by the English Language Subject Benchmark Committee (ELSBC) that was set up to make recommendations to the HKSAR Government on language benchmarks for teachers. One of their major decisions was to deliberately and strongly recommend that the assessment of classroom language should take place in live classroom settings. Since no example of such procedures being carried out in other forms of teacher certification (apart from full-time professional courses run at universities for postgraduate diplomas and the RSA/UCLES teacher certificates and diplomas) could be found at the time, washback considerations were a major issue—particularly in terms of the huge costs and logistic requirements required to carry out the assessments. The washback effect was, of course, instantaneous, with course providers for pre-service language teachers and in-service course developers immediately building into their language benchmark programmes components on classroom language awareness andpractice.

The Role of Stakeholders

In the process of high-stakes assessment, there are normally a number of stakeholders, all with varying degrees of involvement in the process. Clearly, the participants in any form of benchmarking include those who initiate the process, those being assessed, those implementing the process, those who do the assessing and those who certify the process. It is possible and, indeed, likely that none of those assigned to these categories are involved in the process more than once, i.e., those being benchmarked are unlikely to be assessors of benchmarks, developers of benchmarks or implementers of benchmarks. Each role has separate and distinct functions. Other stakeholders may be trade union officials whose role may be to support and defend their members rather than to seek to be involved in setting standards, particularly if it is considered that some of its members may not reach the standards that have been set. Such a role is an uncomfortable one. Trade unions may wish to behave professionally but have to act as a defender of jobs, even though they recognise that not all their members are likely to reach set, agreed benchmarks. The role of government in maintaining high levels of information, education and the dissemination of arrangements for the implementation of benchmarks is crucial. The inclusion and engagement of as many stakeholders as possible in a benchmarking project are usually seen as a vital ingredient for the overall success of the benchmarking project.

Methodology for the Investigation—and Data Collection

The methodology used in the collection of data for any investigation is affected by the philosophical stance adopted by the researchers. Setting benchmarks requires the collection of evidence that can be analysed and interpreted so that, eventually, enough data is collected upon which to base the benchmarks. In order to create rich, 'thick', data during the investigative phase, it is important to use as many data sources as possible (Lincoln & Guba, 1985; Denzin & Lincoln, 2011) if they are to provide the data required to formulate benchmarks.

However, in order to begin to develop constructs for criterion-referenced benchmarks, the first form of data collection should be the sampling of the on-task performance of the clients being investigated. When it is not dangerous to carry out sampling of performances, e.g. when setting benchmarks for language teachers (or for any other cohort of professionals), and little or no disturbance is created by the collection of data, it is the investigators' first priority to collect data in the clients' workplace.

In the Hong Kong case study of English language teachers, described in Chap. 6, the observation of classroom language made it possible to collect, transcribe and analyse the data from which the constructs which underpin teacher language could be developed. Subsequently, descriptors for the four constructs that had been identified were created. In addition, interviews with and observation of teachers led to other constructs (those required for the professional life of a teacher) being identified and, later, assessed.

Authenticity of Task

In high-stakes forms of assessment, test takers find it difficult to accept any form of assessment which is not, at first glance, relevant to the work they do either directly or indirectly. Bachman and Palmer (1996) defined 'authenticity' as the degree that test task characteristics correspond to those in Target Language Use situations. The resistance to some forms of high-stakes testing of teachers in the USA (see, e.g., the case of Massachusetts in the USA—Madaus, 1988) was fuelled by the perception that multiple-choice tests were not the best way to test a teacher's knowledge, understanding and practice of educational principles. Authenticity has been one of the key issues addressed in language tests in recent years [see, e.g., discussion over the communicative language testing that emphasised real-life tasks and authenticity, and performance (Fulcher, 2000) and formats and model of delivery of listening tests (Taylor, 2012)]

As will be illustrated, test takers in the case study found the Classroom Language Assessment the most relevant form of assessment. They also eventually perceived the other performance tests, viz Speaking and Writing, to be authentic and linked to tasks that teachers of English have to perform. After detailed explanations of and experience of taking the other forms of assessment, they also felt that the reading tests were appropriate and relevant although problems with the Listening Test persisted for some time.

Ethics

The issue of *ethics* has always existed in high profile fields such as medicine (e.g. the role of fertility clinics, cloning and the use of brain cells and stem cells in creating life forms). However, the use of the term *ethics* is now being used regularly in academic life (the use of animals in experiments and the use of human 'subjects' now sometimes referred to as 'data points' in research).

As early as 1972, the National Council on Measurement in Education (NCME), the Association for Measurement and Evaluation in Guidance (AMEG), and the American Association for Counselling and Development (AACD is now known as the American Counselling Association) developed a position paper on the responsible use of tests that was intended to ensure that tests are given, and examinees are treated, fairly and wisely (AMEG, 1972). Later in the 1970s, AACD developed a statement on the responsibilities of the users of standardised tests, a document that was revised in 1989 (AACD, 1989). Ethical issues in assessment entered the research literature in 1972 (Schmeiser, 1995 refers to the decisions outlined in the above paragraph). Researchers such as Hamp-Lyons and Lumley (2000), and Bailey and Butler (2004) also discuss issues such as participant involvement in assessment, the test taker's right to the release of results, issues of test taker privacy, test taker rights in the pretesting of forms of assessment, confidentiality, disclosure and anonymity. The use of indirect testing to make predictions about test takers in high-stakes assessments was beginning to be questioned at this time, hence the publication of the American Educational Research Association's (AERA) guidelines, the *Position Statement Concerning High-Stakes Testing in PreK-12 Education* (2000). The issue of ethics has also been addressed regarding accommodating test papers to cater for the needs of candidates in certain minority groups or with special needs, for example, visual, hearing or other physical impairments. However, careful consideration must be given to changes as such changes in test context, format and delivery may change the construct and inferences that can be made from the score (Taylor & Angelis, 2008).

One difficult area in ethics is the production of exemplars in high-stakes assessment procedures that contain performances by participants. This occurs when exemplar material is required for presentation purposes. In the production of video-recorded samples which show test takers taking the test it is ethically unfair to show test takers taking the test to others without first gaining the test takers' approval and indicating to them the audiences who will watch them taking the test. Prior permission must be obtained.

Transparency (Including the Need to Publish)

High-stakes examinations, fraught as they are with tension, can only benefit from attempts to make them transparent. If it is clear to the potential test taker what the benchmark is, what it consists of, what exemplars exist and whether they are easily publically available, what marking schemes are being used (made more transparent by the use of criterion-referenced assessment with its accompanying scales and descriptors), levels of anxiety are likely to decrease. As the UK Academy of Medical Royal Colleges put it (2015:7):

> Since no single method and no single set of procedures can guarantee the defensibility of the standard, there is a duty of transparency towards all stakeholders around the various decisions and their implementations. Documenting how due process was followed allows the stakeholders to see the systematicity of the approach, and therefore forms part of the defensibility evidence for the standard. Following due process may at times result in uncomfortable outcomes, such as a 0% pass rate, or a different pass mark on different days of an examination. Transparency and clear communication about the process should help maintain both good practice and the acceptability of its outcomes to all stakeholders.

Part of the notion of transparency is the willingness of the 'paymaster'/the client to allow findings of ongoing investigations into high-stakes examinations to be published and disseminated. The more that can be added to the public domain the higher the level of transparency of the assessment being considered. The authors were grateful to the HKSAR Government and the HKEAA for allowing them to publish the findings of the investigations they carried out into the validity of the LPATE.

Time Frames—Lead-in Periods

One of the major issues in language benchmarking is the issue of lead-in time in the formulation, preparation and implementation of a battery of assessment instruments. Inevitably, there will be a tension between the time frame that the client wants and the time frame that the researchers and assessment developers feel is required in order to do the job well. The test of who has won in this struggle is the amount of time deviation from/adherence to normal practice in the development of the battery.

Issues Involving the Mixing of Criterion-Referenced Assessment and Analytically Marked Tests

One assumption, accepted by test developers worldwide is that in order to create a battery of tests large enough to satisfy the demands of a high-stakes assessment mechanism, it might be necessary to develop a mixture of tests and test types. A major issue arises when the assessment procedures consist of a mixture of criterion-referenced assessment procedures and tests that are analytically marked.

The issue becomes one of how to calibrate analytically marked tests (such as tests of reading and listening) with criterion-referenced assessments. Criterion-referenced assessment enables a test taker profile to be created where the grades/standard/benchmarks which have been achieved by the test taker can be described on the certificate or assessment report form.

Traditional forms of reading and writing have been used for many years for purposes of norm-referenced assessment. In such cases, it does not matter that one test may be more difficult or less difficult than another because each time the test is administered, it is administered to a similar whole-population cohort and is used for selection or promotion purposes because it ranks the test takers. Such a process does not match the requirements of a benchmark test because a benchmark test wants only a cut score.

However, the problem of what the ‘cut’ scores should be still has to be faced. A ‘cut’ score is required for analytically marked tests in a battery of tests which also includes criterion-referenced tests. There are a number of methods that can be used but basically they come down to two major approaches. The first is the use of expert judges using either the Nedelsky method or the Angoff method. The essence of this approach is that the judges (at least 10–20 in number) make decisions about each item in the question paper and decide whether or not a borderline-pass test taker would score/pass on that item. The sum total of these scores are then added together and divided by the number of assessors. The figure that is reached by these means becomes the ‘cut’ score. This issue is addressed in much greater detail by Drave in Section III of this volume.

The other major method is to choose the criterion-referenced test in the battery which best fits the benchmarks, e.g. the Classroom Language Assessment in the LPATE case study. The grades awarded on that benchmark are then used as a basis to statistically analyse the analytically marked tests using Rasch measurement techniques. The cut scores for the analytically marked tests that are produced by this method are used as benchmarks, for example reading and listening tests.

Exemptions

This is always a contentious issue when benchmarks are being set. A normal response in industry, when dealing with materials, is that materials affected by the new (or upgraded) benchmark must conform to benchmark standards from an agreed date. When personnel are affected, time is normally allowed for existing staff to be upgraded through development programmes or for new staff to be recruited. In certain cases, when it can be shown that certain categories of personnel already meet the new or upgraded benchmarks, exemptions are permitted either on a category or case-by-case basis.

Formal Tests or Continuous Assessment?

Linked to the issue of exemption is the issue of whether to use a one-off form of benchmarking assessment through a battery of assessment instruments at designated intervals or to carry out continuous assessment over time to discover whether participants eventually meet the benchmarks. There are arguments for both types of assessment. When the benchmark involves personnel, a one-off set of assessments can accomplish a great deal quickly. It can also be used diagnostically to indicate whether and in what areas staff may require assistance in order to attain the benchmarks that they have ‘failed’.

Issues Pertaining to the Case Study

A considerable amount of money (US$ 30 million) was set aside by the HKSAR Government to allow teachers to attend development and immersion courses in order to try to attain the benchmark.

Within the context of teacher language assessment, an important issue is whether language proficiency can be divorced from knowledge and awareness of language (subject content) and the ability to use appropriate teaching materials and resources at an appropriate level for students (pedagogic content knowledge). These issues are addressed in the case study, particularly in the Writing Test (Tasks 2a where student language errors had to be corrected and 2b where student language errors had to be explained) and the Speaking Test (Task 3 where three test takers had to discuss a student composition).

Summary

Chapter 4 describes the background to the education system in Hong Kong, and Chap. 5 describes the methodological approaches used in the benchmark case study. Chaps. 6–9 trace the history of the benchmark initiative from its origins in 1995–1996 to its validation and implementation by the HKSAR Government in 2000–2001. The remainder of Section I therefore contains six chapters, as follows:

Chapter 4	Date	An overview of the Hong Kong education and examination systems
Chapter 5	1996	An account of the study's methodology and various statistical techniques and software packages used in Chaps. 6–9
Chapter 6	1996	The initial consultancy feasibility study
Chapter 7	1997–1998	Validation studies and the work of the English Language Benchmark Subject Committee
Chapter 8	1999	The Pilot Benchmark Assessment (English) test bed study, the PBAE
Chapter 9	2000	Determining benchmarks after the PBAE

References

Academy of Medical Royal Colleges (AMRC). (2015). *Guidance for standard setting: A framework for high-stakes postgraduate competency-based examinations*. London: UK. Retrieved December 2016, from http://www.aomrc.org.uk/publications/reports-guidance/standard-setting-framework-postgrad-exams-1015/.

American Association for Counselling and Development (AACD). (1989). *The responsibilities of users of standardized tests. AACD/AMECD policy statement: The RUST statement revised*. Retrieved January, 2018 from http://aac.ncat.edu/Resources/documents/RUST2003%20v11%20Final.pdf.

American Educational Research Association (AERA). (2000). *Position statement concerning high-stakes testing in Pre K-12 education*. Retrieved January, 2018 from http://www.aera.net/About-AERA/Position-Statements.

Association for Measurement and Evaluation in Guidance (AMEG). (1972). The responsible use of tests: A position paper of AMEG, APGA and NCME. *Measurement and Evaluation in Guidance, 4*(2), 385–388.

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

Bailey, A. L., & Butler, F. A. (2004). Ethical considerations in the assessment of the language and content knowledge of US school-age English learners. *Language Assessment Quarterly, 1,* 2–3.

Cheng, L., & Curtis, A. (2012). Test impact and washback: Implications for teaching and learning. In C. Coombe, B. O'Sullivan, P. Davidson, & S. Stoynoff (Eds.), *Cambridge guide to second language assessment* (pp. 89–95). Cambridge: Cambridge University Press.

Denzin, N. K., & Lincoln, Y. (Eds.). (2011). *The SAGE handbook of qualitative research*. CA: SAGE Publications Inc.

Fulcher, G. (2000). The "communicative" legacy in language testing. *System, 28*(4), 483–497.

Hamp-Lyons, L., & Lumley, T. (2000). *Ethical dilemmas in language testing: What can we actually do?* Paper presented at the 22nd Annual Language Testing Research Colloquium, Vancouver, Canada.

Lincoln, Y., & Guba, E. (Eds.). (1985). *Naturalistic inquiry*. Newbury Park, CA: SAGE Publications Inc.

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education, 65,* 29–46.

Schmeiser, C. B. (1995). *Ethics in assessment*. Greensboro NC: ERIC Clearinghouse on Counseling and Student Services.

Taylor, L. (2012). Ethics in language assessment. In C. A. Chapelle (Ed.) *The encyclopedia of applied linguistics*. https://doi.org/10.1002/9781405198431.wbeal0393.

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.

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.