Dealing with the Demands of Language Testing and Assessment

Book Section

Dealing with the demands of language testing and assessment

Glenn Fulcher and Nathaniel Owen

Introduction

A remarkable amount of teacher time is devoted to assessment and testing. Assessment is the broader term, encompassing any activity in which data is collected from learners from which we make judgments about their proficiency or progress. It includes, for example, informal classroom quizzes, and peer- and self-assessment, and is free from the restrictions imposed by formal testing (Fulcher, 2010: 67-92). Testing is the more specific term that refers to formal or standardized testing for the purposes of certification or decision-making. Tests are usually constructed and provided by examination agencies with the authority to issue certificates that are recognised for purposes such as entry to higher education, or employment.

The purpose of this chapter is to introduce teachers and students of applied linguistics to the key concepts and terminology needed to understand the role of standardized language testing and of assessment in the language classroom, and to utilise the testing and assessment literature. We discuss assessment for learning, i.e. how teachers can embed assessment into classroom activities to enhance learning opportunities. We consider the politics of externally mandated testing, and the role of teachers in preparing learners to take examinations. This is a highly controversial area, and so we make recommendations for best practice based upon the concept of 'reverse engineering'. The chapter concludes with a discussion of assessment literacy for language teachers.

Critical issues and topics

Motivation and Learning

Why do teachers spend so much time preparing learners to take tests, and writing tests for their own use? Teachers often hold two related beliefs about the value of tests. The first is that a test acts as a motivational tool for study and learning, incentivising increased effort on the part of the learner (Eklöf, 2010). This is a long-held belief. Ruch (1924) referred to a test as 'the day of reckoning', which encourages learning through establishing clear intermediate and long term

goals. It is also recognized that working towards a qualification is a source of motivation that may not otherwise be present. As Latham (1877: 146) recalls from the introduction of the very first school assessments, 'The efficacy of examinations as a means of calling out the interest of a pupil and directing it into the desired channels was soon recognized by teachers.'

The second belief is somewhat more contentious. In a nutshell, the view is that if tests are designed to carefully reflect the knowledge, skills and abilities (KSAs) that are the goals of an educational programme, the existence of the test will guide the efficient and effective learning of relevant content as well as motivate learning. This is usually termed *measurement driven instruction*. Popham (1987: 682) expresses the strongest form of this position: 'It can be empirically demonstrated that a carrot-laden cart will provide a powerful incentive if placed in the path of an underfed pony'. The argument is clear: it is the test that creates a challenge.

Classroom Based Assessment

Classroom Based Assessment (CBA) is an overarching term for the use of assessment to promote learning, and includes approaches such as Assessment for Learning (AFL) (Turner, 2012). Classroom assessments are designed to provide feedback that leads to learning. Swain (2000: 100) refers to this as the 'notice the gap principle'. Feedback is *descriptive* rather than *evaluative* (Rea-Dickins, 2006). The purpose is to help the learner to become aware of the current state of their L2 development, and what needs to be learned to reach the desired target. Research suggests that appropriate classroom assessment practices lead to improved learner self-esteem and motivation, lower needs for extrinsic reward, and raised levels of success (Black et al., 2003). It therefore relates directly to learner motivation (see Lamb, this volume). Classroom based assessment activities are flexible and open, including the creation of portfolios, group discussions and presentations, process writing, and multi-stage simulations. The guiding principle for the creation of activities is the engagement of learners in communication. For this reason group work is highly valued. Similarly, the assessment of learning outcomes is not limited to the teacher. Self- and peer-assessment are highly valued, which requires learner training in the recognition of their current abilities, in comparison with where they wish to reach. The role of the teacher is one of facilitation, and key teaching skills include effective questioning and providing quality feedback. The latter may include verbal or written information, but scores are not required. Much more important is ensuring classroom

time for learners to respond to feedback. Managing effective classroom assessment requires exceptional classroom management and lesson planning skills.

External Tests

The reason for using external tests has been termed the *test mandate* (Davidson and Lynch, 2002). The mandate may be under the control of schools, as they select tests that best meet the certification needs of their learners. At other times, the test is imposed by an education authority in order to implement a national syllabus (measurement-driven education policy), or to introduce accountability practices for teachers and schools. The use of test data to construct school league tables, or to rank order countries according to scores on language tests, creates tools by which teachers and institutions are evaluated. Evaluations may be directly linked to interventions in the school, disciplinary measures, or teacher pay (Mansell, 2007). Shohamy (2001) identifies this use of language tests as the primary reason for teachers' dislike and mistrust of testing. She argues that it enforces the values of the elite, and punishes those who do not conform. While it is true that testing can have negative impacts such as social exclusion, history shows that in some cases it has also spurred curriculum development, and been instrumental in achieving significant advances for the underprivileged and disenfranchised (Mill, 1859/1998: 118–119). For example, without the evolution of public examinations in the 19th century, the goals of opening the professions and civil service to the middle classes on meritocratic principles would not have been possible (Roach, 1971). The establishment of examination boards such as the Cambridge Syndicate was part of this reform movement (Watts, 2008: 40–41). The use of testing in social reform has arguably shaped the provision of unbiased access to educational and employment opportunities in societies across the world (Zeng, 1999). It is therefore possible to make a positive case for the role of testing as a tool for the maintenance of meritocratic processes in modern democracies (Fulcher, 2015).

Washback

The use of external tests has been shown to affect the work of teachers in schools, and the expectations of learners. The study of testing effects is termed *washback* (Wall, 2012). Alderson and Wall (1993) developed a set of 'washback hypotheses'. These include: a test will influence teaching; a test will influence learning; a test will influence what teachers teach; a test will influence how teachers teach; a test will influence the rate and sequence of teaching,

and so on. While it has been discovered that washback is endemic in education, it has been impossible to articulate a theory that predicts where or how washback will occur. Nor are we much closer to isolating contextual variables that are predictive of how teachers or learners will react to the use of tests. For example, under what conditions might a teacher revert to the use of 'past papers' as teaching materials, linked to a strategy of classroom test-taking followed by analysis of correct/incorrect answers? Concern over the impact of tests on education has also resulted in a new focus on test development. If new tests can be designed in such a way that they promote positive washback, we may be able to avoid the worst consequences of test-preparation (or cramming) that focuses only on test-taking strategies rather than language learning. As Fulcher and Davidson (2007: 144) argue, 'The task for the ethical language tester is to look into the future, to picture the effect the test is intended to have, and to structure the test development to achieve that effect. This is what we refer to as '*effect-driven testing*'.

Key concepts

Test Purpose

Testing always has a purpose. Carroll (1961: 314) expressed this most eloquently: 'The purpose of language testing is always to render information to aid in making intelligent decisions about possible courses of action.' A test is constructed of one or more *prompts*, to which learners respond. The responses are treated as *evidence* for the existence some knowledge, skill or ability (KSA), and the degree of its presence. In classroom based tests, the evidence is treated *formatively*, which means that the evidence is used to enable learners to see where they are at present, and how they may improve. In external proficiency tests, the evidence is treated *summatively*, which means that decisions are being made about an individual at the end of some programme of study. Other purposes for testing include placing learners in suitable classes so that learning materials are challenging but not too difficult, or assessing their achievement against a set of learning objectives over a specified time.

Stakes

The consequences of test outcomes may be fairly benign, as is always the case with classroom based assessment. These are low-stakes assessments. However, in many cases the consequences are high-stakes. This is usually the case with external proficiency tests, the

outcome of which may mean that the test taker cannot, for example, attend university, may not graduate, cannot apply for a job, or move to another country. Teachers must understand whether an assessment or test is high stakes for their learners, as this is likely to impact upon their motivation and the kinds of activities or tasks that they wish to focus on in class.

Referencing

Tests are usually classified as either norm-referenced (NR) or criterion-referenced (CR). A NR test is so called because it is designed to create the maximum discrimination between test takers along a continuous scale. The main requirement for interpretation is the normal distribution of scores. The meaning of the score is its place in the continuum, because the test taker is being compared with all other test takers drawn from the same population. This may be useful when it is necessary to select a predetermined number of candidates, for example. A CR test focuses on a distinction between *mastery* and *non-mastery*, with reference to some absolute definition that resides in a real-world domain (Fulcher and Svalberg, 2013). CR test interpretation asks the question whether a learner has achieved mastery to perform as, for example, an air traffic controller, or an international teaching assistant, irrespective of the scores of other test takers.

Validity

In high-stakes testing, validity is about whether and to what extent our inferences about the meaning of a test score are true. A validity argument sets out the evidence and theory to support the claim we make for score meaning. The kinds of evidence that we would typically expect to find in an argument may include a comparison of the test content with the kinds of communication we would find in the real world. We may find studies relating test scores to an external criterion, such as academic performance in the first year of university study. In some cases, the differential performance of known groups of *masters* and *non-masters* may show that the test is capable of discriminating between target groups. For example, we may find evidence from conversation analytic studies that show learner speech reflects a predicted range of functions or conversational features. Thus, the type of evidence required to support a validity argument depends upon the claims we make for the scores. The theory we expect to see in the validity argument provides the rationale for claiming that the evidence presented supports the proposed score interpretation. In low-stakes tests, particularly classroom based assessments, validation evidence may be collected informally. Validity questions we might ask would

include: Has the feedback resulted in learner improvement? Are the tasks engaging and challenging for learners at this level? Is learner motivation improving?

Reliability and Dependability

In high-stakes tests it is essential to demonstrate that scores are *reliable*, which means that they are *consistent* across facets of the testing context. Such facets usually include time, place, interlocutor, and rater/marker. If a test assesses proficiency to engage successfully in service encounters, the score should not change if the test is taken two or three times over one week, apart from normal random error due to chance factors; over longer periods of time a score may of course change, either because of further study or attrition. Similarly, the score should not vary depending upon where the test is taken, who the interlocutor may be, or who rates the performance, asthese facets are irrelevant to what we wish to assess.

It should be noted that any claim regarding "irrelevant facets" is directly related to the theory underlying the test. For example, if it is claimed that speaking is a socially co-constructed activity, and that the demonstrated proficiency of an individual is variable depending upon the speaking partner, the interlocutor facet suddenly becomes relevant. In this case, it would be essential to collect a number of speaking samples holding all facets of the task stable, but using a different interlocutor for each sample.

The concept of *dependability* is similar to that of reliability, but applies to CR assessments. The question in this case is the extent to which individuals are classified in the same way across facets. It is not just a matter of obtaining reliable scores, but of making dependable classifications of individuals to meaningful categories. Within educational institutions, for example, it would be highly desirable if any individual learner was classified as a master or non-master irrespective of the member of staff conducting the assessment, or the particular task selected. If we are able to do this, our decision making processes are strengthened, and outcomes are more easily defended.

Reliability and dependability are important because they reflect concerns for *fairness* in all assessment practices. The over-riding principle is that all learners should have an equal opportunity at the point of assessment. In practice, this implies that there should be no *bias* towards or against any subgroups of the population, and that the scores should be *independent*

of test method facets that are irrelevant to what we are assessing. In high-stakes contexts it is also important to be aware that a failure to maintain either principle may lead to charges of discrimination against test takers (Fulcher, 2014).

Key areas of dispute and debate

Policy, Politics and Fairness

We have already intimated that external testing increasingly impinges upon language learners and teachers. There is a vocal critical applied linguistic voice arguing that 'Test results have detrimental effects for test takers' (Shohamy, 2001; 15) because the very purpose is to *discriminate*. In testing, discrimination is usually considered to be a 'good thing', because it separates test-takers out for decision making purposes. However, for critical applied linguists *discrimination* retains its usual pejorative meaning.

Similarly, there is a view that all washback on teaching and teachers is negative. Smith (1991) has summarized the arguments that are now periodically rehearsed in the literature. When used for accountability purposes, test scores are published by school or region, and this can lead to negative morale among teachers whose students do not score highly. This potential shift in the use of test scores from the assessment of individual learning to reporting average group performance can be used to evaluate teachers and educational institutions. In turn, this puts pressure on teachers to spend more time attempting to raise average test scores, rather than delivering a broad curriculum. This is precisely the opposite of Popham's claim (noted earlier in the chapter) that measurement-driven instruction improves learning. Many externally mandated tests also use large numbers of multiple-choice items. There is a very good reason for this. The multiple-choice item raises the reliability of any test because it allows the collection of more individual pieces of evidence about performance, and it is possible to create items that discriminate exceptionally well for a known population. However, many teachers use multiple-choice items for teaching and examination preparation at the expense of creative language learning activities.

Standards

Accountability is facilitated by the growth in standards-based assessment, which is driven by the national or international use of *standards*, to which testing practices are obliged to conform (Hudson, 2012). Standards documents are usually made up of subject content with sets of arbitrary levels against which learners are matched (Cizek and Bunch: 18). In language testing, the most commonly used standards documents are those of the American Council on Teaching English as a Foreign Language (ACTFL), the Canadian Language Benchmarks (CLB), and the Common European Framework of Reference (CEFR) in Europe. Their use may be fairly benign, providing a framework for the development of assessments for particular purposes; or they may be used for the implementation of policy and the harmonization of educational systems (Fulcher, 2004). In worse case scenarios, teachers are required to align their tests to the standards, thus removing a great deal of professional discretion from classroom assessment. This could be interpreted as the 'invasion' of the classroom by standardized testing practices. Crease (2011) refers to such a super-system as a *metroscape*. It represents an attempt to ensure that the outcomes of all tests and assessments can be quantified in the same 'currency', which is given its value by the institution that prints the units and recognizes them for use in education and employment. The effect has been the creation of an *alignment industry*, which has come to view mapping test scores to external standards as a validation process in its own right. For example the mapping of institutional tests to the CEFR (e.g. Martyniuk, 2010). There are always (at least) two sides to every case. For some, the use of a common currency makes sense of qualifications and certificates are to be accepted across national borders (Jones, 2013). For others it is a feature of unwanted political control that undermines the use of tests for defined purposes, subverts validity theory, and threatens teacher independence (Fulcher, 2016).

Implications and challenges for ELT practice and practitioners

Living in the modern Metroscape

The first challenge for ELT practitioners is to work out just where they stand on the contentious issues that we have outlined. Do you believe in measurement-driven teaching, or does the dominance of tests undermine your professionalism and the quality of the learning experience? Should you use tests to motivate learners, or aim to develop intrinsic motivation for learning? Do you wish all your tests to be based on, and interpreted in, terms of an external set of standards? Or do you believe in ecologically sensitive assessment where interpretation is

context dependent? The answer to these questions does not come entirely from assessment theory, but from the values that you bring to the profession of language teaching.

In this section, we wish to illustrate the implications of value-driven choices in one particular area. We have already suggested that the use of external tests has a profound effect on teacher and learner behaviour. We know this from washback research. Learner motivation to pass external tests also brings pressure on teachers to engage in test-preparation practices. The metroscape also enlists parents or sponsors, principals and heads, to create an environment in which the teacher is required to 'get students through the exam', and upon which their effectiveness is judged. This is a familiar scenario in language education the world over. We believe that teachers faced with these pressures can use classroom based assessment practices to deal with the demands of external testing regimes.

Test Preparation Practices

Popham (1991: 13) suggests that classroom practice should be guided by two principles. The first is the educational correlate of the Hippocratic Oath, that 'no test preparation practice should violate the standards of the education profession.' This essentially rules out raising test scores by altering them during or after the test, or assisting others to circumvent test security. It also prevents teachers from excluding learners expected to get low scores from taking the test in order to artificially inflate averages for the benefit of the institution's place in league tables. The second is the educational defensibility clause, which states that 'No test preparation practice should increase students' test scores without simultaneously increasing student mastery of the content domain tested.' This principle rules out focusing on test-taking strategies, such as learning tricks for guessing the correct answer in a multiple-choice item, or memorizing canned responses to speaking or writing prompts (pre-prepared answers to lists of common questions). Haladyna et al. (1991) refer to the score raising effect of these practices as 'score pollution'. In such cases, the inference from score to what a learner can do is weakened or destroyed completely. Validity is undermined. In very high-stakes contexts such as air traffic control, the results of unethical practice can also be extremely dangerous.

The question we therefore wish to address is how language professionals might ethically prepare learners for tests? This question is specifically considered with the impact of external tests upon teachers, and we choose this as a focus because we think it is one of the greatest

challenges that teachers face in the metroscape that is set to dominate language education for the foreseeable future. What we wish to do is offer an approach to test preparation that maintains the principles set out by Popham.

Reverse Engineering

One solution that we propose is for language teachers is *reverse engineering* (RE). This is defined as 'the creation of a test specification from representative sample items' (Davidson and Lynch, 2002: 41). A *test specification* is the blueprint, or design document, that states what the test is designed to assess, and how it should do that. A test specification therefore contains a statement of the purpose of the test, and the range of items or tasks that the test should contain. It may also include information about the topics that might be used, where texts should come from, and how difficult they might be. It also contains sample items or tasks, so that someone who is asked to write them could see what was intended by the test designer. Specifications are never put into the public domain. What teachers and learners see is a general description of the test, along with sample 'papers' that illustrate what might be expected to appear on any particular test form.

Reverse engineering is the process of analysing the content of the test in order to recreate the specification. The reverse engineered specification does not have to be exact; it merely has to be accurate enough to identify the key skills, ability or knowledge that the test was designed to assess. The process may be aided by any documentation that the test providers do place into the public domain for test takers and teachers.

Fulcher and Davidson (2007: 57–58) identify test deconstruction reverse engineering as relevant to teachers who wish to prepare learners to take tests. The purpose of this type of RE, carried out by teachers, is to discover what the test is designed to assess, uncover the KSAs that the test writers value, and relate these to the goals of learners in the institution. The intention is not to produce more items or tasks similar to those in the test, but to design learning activities suitable for a classroom that target the skills valued by the assessment system. When teaching reading, for example, an analysis of the variety of genres used, topics, lexical range, difficulty, and target audience, may inform the search for learner reading material. The item types will reveal the reading skills valued, such as understanding explicit information, identifying main points, summarising detail, skimming, scanning, interpreting diagrammatic

information, making inferences, or comparing and contrasting different arguments across texts. Such an analysis provides a rich starting point for the design and production of creative learning activities that do not rely on testing practice or test-type items. It is hypothesised that the use of learning activities that are much closer to the kind of tasks found in classroom based assessment are more likely to lead to learning, than the use of the kinds of tasks found in tests. If this is the case, as the classroom based assessment research claims, learners will acquire the target skills without engaging in test taking strategies. Such learning, it is claimed, will translate into higher test scores if the tests are valid measures of those skills.

Examples of RE to support ethical test preparation are provided by Owen (2011). Analysing one item in relation to an IELTS reading text on population movement and genetics provides rich information to inform reading pedagogy; we reproduce two paragraphs from the text and the test item in Figure 1, and its RE deconstruction in Figure 2.

[Insert Figure 1 here]
[Insert Figure 2 directly following Figure 1]

The first important observation is that no background knowledge is required to answer this item correctly. This is provided by the glossary, which may be substituted in a classroom environment by pre-teaching lexical items. The selection of the correct response then involves a three-stage process. The first is the identification of the paragraph that contains the required information, and this can be done by scanning to match lexical items from the prompt to text. The selection of answer C requires an ability to understand and identify the common textual structuring device of enumeration. Once identified, the Inuit group must be matched to the correct wave by recognizing synonymy between prompt and text.

Armed with this analysis, the teacher may design reading tasks drawing upon texts of varying and increasing complexity, that assist learners in scanning, identifying discourse structure markers, and recognizing synonyms/antonyms. This is good practice in reading pedagogy, not mindless test preparation (Grabe and Stoller, 2013).

Future directions

The language classroom has become increasingly complex. As language teachers we are required to reconcile innumerable conflicting priorities and demands. We believe that the modern metroscape is perhaps the most significant factor that impinges upon classroom practice. This is perhaps not surprising - test scores have become a commodity. They are the vehicle by which learners travel on to further educational opportunities, or acquire positions that will secure their economic future. The desire to get higher test scores can even become an end in its own right, which is why some learners begin to believe that short cuts (such as cheating) are a viable option. We are not going to change the market value of test scores. The challenge facing us is therefore engaging with learners, their sponsors, and the institutions for which we work, to show that sound pedagogy and ethical practice are not in conflict with success in tests.

The example of Reverse Engineering that we have given is a strategy to identify what is valued by high-stakes tests, and to create non-test tasks that target those skills and abilities. These tasks may be used in classroom based assessment for learning. This begins to break down the division between the summative and the formative. It does assume, of course, that the external test has been designed to assess abilities that we wish to teach, and which are relevant to the future success of our learners. For the most part, the examination boards and agencies that create high-stakes tests do have a theoretical basis that links test content to a defined purpose. But if we believe and can show that this is not the case, we have a responsibility to look for alternative tests that will support rather than frustrate our pedagogy.

This is an ambition to keep the responsibility for teaching and learning firmly with the classroom teacher. To deal with the challenges we have outlined, it is essential that all language professionals are assessment literate. In the coming years, one of the key challenges will be to continue to refine our definition of what language professionals need to know about assessment to improve learning (Fulcher, 2012), and devise new strategies for delivering that literacy in both initial teacher education, and continuing professional development. This may be in the form of seminars and conferences. It may be largely delivered online. But best practice will surely be for teachers to work together in local contexts to develop their own learning materials and share best practice in using assessment for learning. Solutions that are ecologically sensitive to specific contexts and learners are almost always preferable to the alternatives. They empower learners and teachers to improve learning, and ultimately, test scores as well.

Discussion questions

- To what extent do tests motivate learners? How might you define 'motivation', and the types of tests that may impact upon it? A related additional reading and a video is available at: http://languagetesting.info/features/motivation/mil.html.
- Examine the principles and methods of Assessment for learning (AFL) available at: http://languagetesting.info/features/afl/formative.html. How might you create the time to implement particular AFL strategies in your own teaching context?
- The Programme for International Student Assessment (PISA) is run by the OECD (Organization for Economic Cooperation and Development). Their literacy tests compare country performance and inferences are drawn about likely future economic performance. Examine the PISA tests and date at their website: http://www.oecd.org/pisa/home/. Are the inferences drawn from scores sound? Are there any flaws in the tests or the steps in the OECD arguments? What are the likely impacts of such testing on education policy in the country where you work? Is there any likely washback on your teaching environment?
- The three most widely used testing standards documents are:
  - ACTFL: http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012
  - CLB: http://www.language.ca/
  - CEFR: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

  Which do you think may be helpful in designing your own assessments? How would you use the document(s) in your own context?
- The International Language Testing Association (ILTA) publishes:
  - a Code of Ethics http://www.iltaonline.com/index.php/en/resources/ilta-code-of-ethics
  - Guidelines for Practice http://www.iltaonline.com/index.php/en/resources/ilta-guidelines-for-practice

  To what extent do you think these offer reasonable guidance for your own classroom practice in test preparation and assessment?

Related topics

Educational perspectives on ELT; Language curriculum design: possibilities and realities; Values in the ELT classroom;

Further reading

The three texts recommended for further reading represent graded reading in language testing and assessment, from an introductory text to an advanced resource book. The advanced book was written first, and the other two commissioned at a later date to provide a coherent progression of topics and difficulty.

Douglas, D. (2010). *Understanding Language Testing*. London: Hodder Education/ Routledge. (An introductory text that explains basic terminology and key concepts in language testing, outlines the skills required to design and use language tests, and introduces simple statistical tools for test analysis. No prior knowledge of language testing is assumed.)

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education/Routledge. (An intermediate text that deals with the purpose of testing in context and an analysis of test use in society. The text then follows the 'test development cycle' to explain in detail the process of test design, implementation, and interpretation.)

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London & New York: Routledge. (The ten sections in this volume each address a key issue in language testing, including: validity, test specifications, the nature of constructs, designing and piloting tests, scoring, ethics and fairness. Each section is accompanied by a key article from the field, and related activities for groups and individuals.)

Website: http://languagetesting.info (An extensive language testing website providing links to articles and journals, videos explaining key concepts, podcasts and features on a range of topics, study scenarios, statistical resources, and daily language testing news from the world's media.)

References

Alderson, J. C. and Wall, D. (1993) 'Does Washback Exist?' *Applied Linguistics*, 14/2. 115 – 129.

Black, P. Harrison, C., Lee, C., Marshall, B. and Wiliam, D. (2003) *Assessment for Learning: Putting it Into Practice*. Buckingham, U.K: Open University Press.

Cambridge University Press (2009) *Cambridge IELTS 7 Self-study Pack: Examination Papers from University of Cambridge ESOL Examinations (IELTS Practice Tests*). Cambridge: Cambridge University Press. 70-73.

Carroll, J. B. (1961) 'Fundamental considerations in testing for English language proficiency of foreign students', reprinted in H.B. Allen and R.N. Campbell (eds.) (1965) *Teaching English as a Second Language: A Book of Readings*. New York: McGraw Hill. 313 – 330.

Cizek, G. J. and Bunch, M. B. (2007) *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. London: Sage.

Crease, R. P. (2011) *World in the Balance. The Historic Quest for an Absolute System of Measurement*. New York and London: W. W. Norton & Company.

Davidson, F. and Lynch, B. K. (2002) *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven and London: Yale University Press.

Eklöf, H. (2010) 'Skill and will: test-taking motivation and assessment quality'. *Assessment in Higher Education: Principles, Policy & Practice*, 17/4. 345 – 356.

Fulcher, G. (2004). 'Deluded by artifices? The Common European Framework and harmonization'. *Language Assessment Quarterly*, 1/4. 235 – 266.

Fulcher, G. (2012) 'Assessment literacy for the language classroom'. *Language Assessment Quarterly*, 9/2. 113 – 132.

Fulcher, G. (2014) 'Language testing in the dock', in A.J. Kunnan (ed.) *The Companion to Language Testing.* London: Wiley-Blackwell. 1553 – 1570

Fulcher, G. (2015). *Re-examining Language Testing: A Philosophical and Social Inquiry*. London and New York: Routledge.

Fulcher, G. (2016). 'Standards and Frameworks', in J. Banerjee and D. Tsangari (eds) *Handbook of Second Language Assessment*. Berlin: DeGruyter Mouton.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London and New York: Routledge.

Fulcher, G. and Davidson, F. (2008) 'Tests in life and learning: A deathly dialogue'. *Educational Philosophy and Theory*, 40/3. 407 – 417.

Fulcher, G. and Svalberg, A. (2013) 'Limited aspects of reality: Frames of reference in language assessment'. *International Journal of English Studies*, 13/2. 1 – 19.

Grabe, W. and Stoller, F. (2013). *Teaching and Researching Reading*. Second Edition. London and New York: Routledge.

Haladyna, M., Nolen, S. B., and Haas, N. (1991)' Raising standardized achievement test scores and the origins of test score pollution'. *Educational Researcher*, 20/5. 2 – 7.

Hudson. T. (2012) 'Standards-based testing', in G. Fulcher and F. Davidson (eds) *The Routledge Handbook of Language Testing*.London and New York: Routledge. 479 – 494.

Hughes, A. (2003). *Testing for Language Teachers. Second Education*. Cambridge: Cambridge University Press.

Jones, N. (2013). 'Defining an inclusive framework for languages', in E.D. Galaczi and C.J. Weir (eds) *Exploring Language Frameworks*. Cambridge: Cambridge University Press. 105–117).

Latham, H. (1877). *On the Action of Examinations Considered as a Means of Selection*. Cambridge: Dighton, Bell and Company. Mansell, W. (2007). *Education by Numbers. The Tyranny of Testing*. London: Politico's Publishing.

Martyniuk, W. (2010). *Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual*. Cambridge: Cambridge University Press.

McNamara, T. and Roever, C. (2006). *Language Testing: The Social Dimension*. London: Blackwell.

Mill, J. S. (1859/1998) 'On Liberty', in J. Gray (ed) *John Stuart Mill's On Liberty and Other Essays*. Oxford, Oxford University Press, 2 - 128.

Owen, N. (2011) *Reverse Engineering: A Model and Case Study*. University of Leicester: Unpublished MA Dissertation.

Popham, W. J. (1987) 'The merits of measurement-driven instruction'. *Phi Delta Kappan*. 68/9. 679-682.

Popham, W. J. (1991) 'Appropriateness of Teachers' Test-Preparation Practices'. *Educational Measurement: Issues and Practice,* 10/4. 12 – 15.

Rea-Dickins, P. (2006) 'Currents and eddies in the discourse of assessment: a learning-focused interpretation'. *International Journal of Applied Linguistics,* 16/2.163 – 188.

Roach, J. (1971) *Public Examinations in England 1850 – 1900*. Cambridge: Cambridge University Press.

Ruch, G. M. (1924) *The Improvement of the Written Examination*. Chicago: Scott, Foresman and Company.

Shepard, L. (2000) 'The Role of Assessment in a Learning Culture'. *Educational Researcher,* 29/7. 4 – 14.

Shohamy, E. (2001) The *Power of Tests: A Critical Perspective on the Uses of Language Tests*. London: Longman/Pearson Education.

Smith, M. L. (1991) 'The effects of external testing on teachers'. *Educational Researcher*, 20/5. 8 – 11.

Swain, M. (2000) 'The output hypothesis and beyond: Mediating acquisition through collaborative dialogue', in J. Lantolf (ed.) *Sociocultural Theory and Second Language Learning*. Oxford: Oxford University Press. 97-114.

Turner, C. E. (2012) 'Classroom Assessment', in G. Fulcher and F. Davidson (eds) *The Routledge Handbook of Language Testing.* London and New York: Routledge. 65-78.

Wall, D. (2012) 'Washback', in G. Fulcher and F. Davidson (eds) *The Routledge Handbook of Language Testing*. London and New York: Routledge. 79-92.

Watts, A. (2008) 'Cambridge Local Examinations 1858 – 1945', in S. Raban (ed.) *Examining the World. A History of the University of Cambridge Local Examinations Syndicate*. Cambridge: Cambridge University Press.

Zeng, K. (1999). *Dragon Gate. Competitive Examinations and their Consequences*. London: Cassell.

A: TEXT

An important project, led by the biological anthropologist Robert Williams, focused on the variants (called Gm allotypes) of one particular protein - immunoglobin G - found in the fluid portion of human blood. All proteins 'drift' or produce variants, over the generations, and members of an interbreeding human population will share a set of such variants. Thus, by comparing the Gm allotypes of two different populations (e.g. two Indian tribes), one can establish their genetic 'distance', which itself can be calibrated to give an indication of the length of time since these populations last interbred.

Williams and his colleagues sampled the blood of over 5,000 American Indians in western North America during a twenty- year period. They found that their Gm allotypes could be divided into two groups, one of which also corresponded to the genetic typing of Central and South American Indians. Other tests showed that the Inuit (or Eskimo) and Aleut* formed a third group. From this evidence it was deduced that there had been three major waves of migration across the Bering Strait. The first, Paleo-lndian, wave more than 15,000 years ago was ancestral to all Central and South American Indians. The second wave, about 14,000 - 12,000 years ago, brought Na-Dene hunters, ancestors of the Navajo and Apache (who only migrated south from Canada about 600 or 700 years ago). The third wave, perhaps 10,000 or 9,000 years ago, saw the migration from North-east Asia of groups ancestral to the modern Eskimo and Aleut.

Glossary for the Whole Text

1 New World: the American continent, as opposed to the so-called Old World of Europe, Asian and Africa

2 modern Native American: an American descended from the groups that were native to America

3 Inuit and Aleut: two of the ethnic groups native to the northern regions of North America (i.e. northern Canada and Greenland)

4 DNA: the substance in which genetic information is stored

5 crown/root: parts of the truth

6 incisor/premolar/molar: kinds of teeth

B: TEST ITEM

The reading passage refers to the three-wave theory of early migration to the Americas. It also suggests in which of these three waves the ancestors of various groups of modern native Americans first reached the continent

Question 22.

Stem: "Classify the groups named in the table below as originating from:

A – the first wave; B – the second wave; C – the third wave.

Inuit 22……. **[ANSWER: C]**

Figure 1: An example of an IELTS test reading item. Source: Cambridge University Press, 2009: 70 – 73.

Textual reference: "The third wave, perhaps 10,000 or 9,000 years ago, saw the migration from North-east Asia of groups ancestral to the modern Eskimo and Aleut."

Response Attribute: test taker should recognize wording 'three-wave theory' from question stem and scan the text to identify the corresponding paragraph (D). Test taker should then recognize textual cohesion throughout the paragraph ("the first…the second…the third…"). Lexical linking of stem words 'originating from' and text word 'ancestral' is required to answer question confidently.

Figure 2: Reverse engineering deconstruction of example test item