1 Wagging the dog? Towards a model of washback

Tests influence teaching and learning. Where a test is used for selection, as is IELTS, those who seek access will attempt to gain the skills they believe necessary to succeed on the test. Some of these skills are generally considered to be desirable, as they are required in the target language use domain. However, as all tests are limited in how much of the domain they can sample and involve a certain amount of measurement error, there is inevitably scope for the misrepresentation of test takers' abilities. The skills required to pass a test are not necessarily or comprehensively the skills required in a target language use domain (Bachman and Palmer 1996). Washback is thus grounded in the relationship between preparation for success on a test and preparation for success beyond the test, in the domain to which the test is intended to generalise and to which it may control access.

This chapter is devoted to discussion of different understandings of washback found in the literature. A conceptual model of washback is outlined that will serve to guide the research described in later chapters. In Chapter 2, the model is applied to the relationship between the IELTS test and theories of academic writing to make predictions about how the test might be expected to influence teachers and learners. These predictions are then tested against the evidence from the research studies described in Chapters 3 to 6.

Washback: definition and scope

In the literature (both in applied linguistics and in general education), the terms *backwash* and *washback*, are both used, and are invariably seen as interchangeable (Pearson 1988, Bachman 1990, Alderson and Wall 1993, Gipps 1994, Broadfoot 1996, Davies et al 1999, Hughes 2003). Dictionary definitions give *backwash* as a backward flow or movement of a fluid produced by a propelling force (as of an oar in rowing) with the meaning extending, figuratively, to the repercussions of a momentous event (Collins 1979, Merriam-Webster 2000, Oxford University Press 2000). Backwash carries technical meanings in fluid mechanics and in the economics of development where economic growth in an urban centre may have beneficial effects on the periphery (*spread*) or conversely, cause decline (*backwash*). As Spolsky

(1996) points out, in general usage the word is most usually applied to unintended and negative effects, and until the 1980s it generally carried a similarly negative meaning within applied linguistics (Khaniya 1990). However, *washback* has gained in currency and is now generally accepted in the applied linguistics literature. The term washback will therefore be used throughout, except where quoting from other writers.

In applied linguistics, washback is broadly defined as the effect of a test on teaching (Richards, Platt and Platt 1992, Davies et al 1999) and often also on learning (Shohamy 1993, Hughes 2003). It has been variously associated with effects on teachers, learners (Buck 1988, Messick 1996, Shohamy 2001), parents (Pearson 1988), administrators, textbook writers (Hughes 1993), instruction (Bachman 1990, Chapelle and Douglas 1993, Weigle 2002), the classroom (Buck 1988), classroom practice (Berry 1994), educational practices and beliefs (Cohen 1994) and curricula (Weigle 2002, Cheng 2005), although for Hughes (1993) and Bailey (1999), the ultimate effects on learning outcomes are of primary concern.

For Shohamy (1992, 1993, 2001) washback is an intentional exercise of power over educational institutions with the objective of controlling the behaviour of teachers and students. For Valette (1967), Wilkinson (1968) and Spolsky (1996) it represents only unforeseen and deleterious effects, while Cheng (1997, 2005) uses the term to refer exclusively to the intended curriculum changes associated with a testing innovation. More commonly, it is considered a neutral term (Alderson and Wall 1993) which may refer both to (intended) positive (Bachman and Palmer 1996, Davies et al 1999) or beneficial (Buck 1988, Hughes 2003) effects and to (unintended) harmful (Buck 1988) or negative effects (Bachman and Palmer 1996, Davies et al 1999, Hughes 2003).

A number of related terms, originating in the general educational measurement literature, have similar meanings and are sometimes equated with washback (Cheng 2005). Among these are *test impact* (Bachman and Palmer 1996, Shohamy 2001), *test influence* (Alderson and Wall 1993); *teaching to the test* (Madaus 1988); *measurement-driven instruction* (Popham 1987); *curriculum alignment* (Smith 1991a, Resnick and Resnick 1992), *systemic validity* (Fredericksen and Collins 1989) and the *consequential aspect of validity* (Messick 1989, 1996). There are often differences of approach implicit in the terminology and it is therefore important to differentiate between washback as operationally defined in this study and other similar terms as they are used in the literature.

Although the terms have been used to refer to the same concept, washback is distinguished from test impact by Bachman and Palmer (1996) who, with McNamara (1996, 2000), Hamp-Lyons (1998) and Shohamy (2001) locate washback under the umbrella of impact. While impact may occur at a *macro* or social and institutional level, washback occurs only at the *micro*

level of the individual participant (primarily teachers and students). Bailey (1996, 1999) makes a further distinction between *washback to the learner* and *washback to the programme*. The former refers to effects on students, while the latter embraces effects on other participants such as teachers, materials writers and administrators.

For Brown and Hudson (2002), Bachman and Palmer (1996) and Bailey (1999) washback is not limited to preparation for taking a test, but may include the effects on an individual of actually sitting the test, of feedback received and of decisions taken on the basis of test scores (Bachman and Palmer 1996). However, these are excluded from other, narrower definitions (Weir 1990, Hamp-Lyons 1991, McNamara 2000), where washback is restricted to effects on the teaching and learning 'prior to', 'preceding' or 'leading up to' a test.

Although the terms have been used synonymously (Peirce 1992, Berry 1994), Shohamy (2001) distinguishes washback from Fredericksen and Collins' (1989) concept of *systemic validity*. Taking a systems approach – viewing test and curriculum as components in a constantly developing educational system – Fredericksen and Collins propose that a systemically valid test will be one that brings about curricular and instructional changes in an educational system that advance the development of the cognitive skills that the test is intended to measure (1989). Washback, for Shohamy (2001), following Alderson and Wall (1993), is more narrowly concerned with teachers and learners, but is equally dependent on comparisons made over time or between systems.

Washback is not generally considered to be a standard for judging the validity of a test. Although Morrow (1986), with support from Weir (1990) and Khaniya (1990), argues for *washback validity*, or the extent to which a test fulfils declared pedagogic aims, as a standard for the evaluation of language test development, Alderson and Wall (1993) and Messick (1996) contest this. Firstly, there are no agreed standards for evaluating washback and individual stakeholders may each regard the same effects differently (Hamp-Lyons 1987, Mehrens 1998). Secondly, washback can only be related to a test indirectly, as effects are realised through the interactions between, *inter alia*, the test, teachers and learners. A well-designed test may therefore be associated with negative consequences because of features of an educational system other than the test (Messick 1996). However, Messick suggests that tests which satisfy validity criteria are more likely to have a positive influence on teaching and learning, and so counsels that washback is not a sign of test validity, but that a valid test is likely to generate positive washback.

On the other hand, Messick (1989) argues that evidence of testing consequences, of which washback is one aspect, needs to be weighed with evidence supporting test inferences in validating test use, a view that has now been widely accepted in the educational measurement community (American Educational Research Association 1999). In this vein, Bachman (1990)

recognises a potential for conflict between washback and other validity considerations so that even though a direct test of writing may have a lower predictive validity than a multiple-choice test in a university entrance test battery, the potential washback on instruction may outweigh this in deciding which test to prioritise (Bachman 1990).

While Popham (1987) views content as primary, holding that testing relevant skills, even through indirect test formats, will encourage those skills to be taught, advocates of communicative testing disagree. They argue that because test preparation involves training in the kinds of activities which appear in a test, maximal authenticity – the degree of correspondence between a given test task and a target language use (TLU) task (Bachman and Palmer 1996) – and directness – the extent to which a test entails a candidate performing precisely the skill(s) we intend to measure (Hughes 2003) – should be fundamental considerations in test design (Morrow 1986, Wesche 1987, Shohamy, Donitsa-Schmidt and Ferman 1996, Hughes 2003). The influence of sociolinguistics and pragmatics on the construct of communicative language ability (Bachman 1990) dictates the integration of skills in meeting test task demands.

Messick (1996) relates directness and authenticity to washback through the general validity criteria of *construct under-representation* and *construct-irrelevant variance*. He demonstrates that construct under-representation threatens authenticity, while construct irrelevant variance threatens directness. A test design which maximises authenticity and directness is therefore held to have the greatest likelihood of fostering positive washback. However, Messick also reminds us that both directness and authenticity are problematic qualities which cannot be fully realised in testing situations (see discussions in Bachman 1990, Lewkowicz 2000).

In construct under-representation (which jeopardises authenticity), a test is too narrow and fails to include important dimensions of the targeted construct (Messick 1996). A test, for both technical and practical reasons, can only address a limited sample of the focal test construct. Some areas of a test construct may not be readily accessible to measurement. Concern for reliability may also dictate a narrow consistency of content and format between testing occasions. As a result, participants are often able to predict test content and may direct their resources to the areas of the construct they anticipate are most likely to be tested at the expense of those that are least likely. Consequently, although users may interpret scores as indicating ability in all areas of the focal construct, the scores may, in fact, reflect a relatively limited knowledge or ability.

In construct-irrelevant variance (which jeopardises directness), the test is too broad, containing features that are not of relevance to the construct (Messick 1996). Examples of construct-irrelevant variance would be test formats that are vulnerable to test-wiseness coaching and so allow increases

in test scores without development of the skills purportedly measured by the tests (resulting in invalidly high scores) or unfamiliar formats that cause confusion for some test takers (resulting in invalidly low scores).

Wiliam (1996) follows Messick (1989) in reversing the concept of washback validity. A test is not valid to the extent that it engenders positive washback, but is likely to engender positive washback to the extent that it is valid. Thus, a test will be valid if one would be satisfied for teachers to teach to the test. However, to be satisfied that teaching to the test would be beneficial, the test would have to meet the criteria for construct validity set out in Messick's (1989) four-faceted validity framework: a) the test would represent the whole of the target domain (within-domain inferences in Messick's framework); b) teaching to the test would also improve learners' performance on other assessments of the same abilities (beyond-domain inferences); c) the test would adequately represent our theoretical understanding of what was important in the domain (within-domain consequences); and d) the effects of teaching towards the test on teachers and students would be beneficial (beyond-domain consequences) (Wiliam 1996).

However, Wiliam recognises that no assessment can achieve such an ideal. Some negative effects are to be anticipated from any test. The relative extent to which different testing technologies are able to overcome the threats of construct under-representation and construct-irrelevant variance is central to discussions of washback and is recognised as a matter for empirical research, contributing to test validation (Linn, Baker and Dunbar 1991, Moss 1992, Alderson and Wall 1993, Shepard 1993).

In brief, most commentators associate washback with test preparation or teaching to the test, excluding broader questions of the social impact of test use and subsequent educational effects on individuals such as those resulting from diagnostic feedback or placement decisions based on test scores, although this limitation is not always made explicit (Valette 1967, Wilkinson 1968, Morrow 1986, Buck 1988, Hughes 1988, 1993, 2003, Pearson 1988, Davies 1990, Heaton 1990, Weir 1990, Watanabe 1992, McNamara 1996, Hamp-Lyons 1998). These constraints on washback studies have been the cause of some dissatisfaction for those language testers interested in exploring wider questions of test consequences (Hamp-Lyons 1997, 1998).

In accepting a relatively narrow definition of washback, exploration of *macro* issues (Bachman and Palmer 1996) of social impact and test use embraced by conceptions of the consequential aspect of test validity (Messick 1994, 1996) are excluded from this study. Also excluded are such *micro* issues as the degree to which test scores guide language support subsequent to a test or how test design might affect students' learning *after* they have completed the IELTS test and progressed to academic study.

Although the IELTS candidature has increased rapidly over recent years with a concomitant rise in the number of preparation textbooks and

preparation courses, the limited period available for the study also largely excludes the important question of how washback from the test has developed over time.

Washback by design: positive washback through authentic, direct testing

Washback is said to vary along at least two dimensions: *direction* (positive or negative) (Buck 1988, Alderson and Wall 1993, Brown and Hudson 2002, Hughes 2003) and *extent* (Bachman and Palmer 1996), *strength* (Gates 1995) or *intensity* (Cheng 2005).

The concept of washback direction encapsulates the principle that some effects of a test may be beneficial to the development of learners' abilities, while others may be damaging. For Bailey (1996), washback can be either positive or negative to the degree that it either encourages or inhibits the attainment of educational goals held by learners, educators or both. Thus washback is often evaluated as positive or negative according to how far it encourages or discourages forms of teaching or learning judged to be appropriate. Of course, what is considered to be appropriate will depend on the position adopted by the judge and the educational goals he or she espouses (Hamp-Lyons 1987, Mehrens 1998).

In reality, learners, teachers, administrators and other participants may have competing goals. Thus effects regarded as positive by one constituency within an educational system may be seen as negative by another: witness for example the impassioned debate over the merits of *measurement driven instruction* in the USA (Bracey 1987, Popham 1987, Airasian 1988, Ramirez 1999).

When tuition is geared to ensuring that students pass a test, at least some effects on teaching and learning are generally acknowledged to be positive. Among the potential benefits of test preparation are enhanced motivation and clearer, more focused instructional targets (Madaus 1988, Alderson and Wall 1993, Gipps 1994). Although some fear that these advantages may be brought through increased anxiety and intimidation (Stiggins 1999), at the least, preparation for a test motivates reflection about the material (Cohen 1994).

Nevertheless, washback has, historically, more often been associated with the negative effects of tests on teaching and learning (Spolsky 1996). Although much of the evidence is anecdotal (Alderson and Wall 1993) and the allegations have sometimes been contradicted by research findings (Wesdorp 1982, Watanabe 1992), the perception of damaging effects is widespread and well-established (Vernon 1956, Wiseman 1961, Cronbach 1963, Kellaghan, Madaus and Airasian 1982, Madaus 1988, Eisemon 1990, Khaniya 1990, Corbett and Wilson 1991, Haladyna, Nolen and Haas 1991, Smith 1991a, 1991b, Kellaghan and Greaney 1992, Gipps 1994, London 1997, Bailey 1999, Jones et al 1999, Shohamy 2001).

Based on perhaps the most thorough study on the subject to date (Bailey 1996, Hamp-Lyons 1998, Cizek 2001), involving a two year qualitative study carried out in US primary schools, Smith (1991b) finds support for the common allegations: testing programmes a) considerably reduce the time available for instruction, b) restrict the range of the curriculum and limit teaching methods, and c) potentially reduce the freedom of teachers to teach content or to use methods that are believed to be incompatible with the format of standardised tests. These findings are echoed in descriptions of testing programmes from payment-by-results in Victorian England (Holmes 1911, cited in Gipps 1994) through the 11-plus selection tests for English grammar schools (Vernon 1956, Broadfoot 1996) to state-mandated accountability testing in the USA (Corbett and Wilson 1991, Hermann and Golan 1993) and testing innovations in developing countries (Eisemon 1990, Chapman and Snyder 2000).

However, evidence from empirical research, as opposed to anecdote, is scanty and such evidence as does exist is not of a quality to support the inference that tests are *responsible* for teaching and learning practices (Mehrens 1998). Nonetheless, the proliferation of preparation courses and coaching materials and the testimony of teachers and students can leave little doubt that at least some tests do influence at least some individuals. It is less clear, however, why this happens and how far test design and use (as opposed to publishers, teacher trainers and school administrators or widely held beliefs about learning, for example) are implicated in any adjustment to behaviour (Hamp-Lyons 1998).

The objective underlying many test preparation practices is to exploit the format and content of a test to improve test scores quickly and efficiently. Apparent success in boosting scores has given rise to concerns that test-preparation activities threaten the interpretability of test scores. Tests do not (and probably cannot) include all of the abilities considered important in a domain (Wiliam 1996). Some skills are easier to test than others and these skills may come to be better represented on tests than other, perhaps equally important skills. Teachers may be able to predict test content and so become able to work efficiently to improve students' test scores by directing their efforts to the tested areas of a curriculum at the expense of untested areas (Crooks 1988).

Test scores are usually interpreted to represent ability in all areas of a domain, but where teaching has focused only on tested skills, such inferences may not be justified. As Cronbach (1963) observes, the knowledge that particular content will be tested encourages a concentration of effort, which is desirable, but learning the answer to a question is not the same as reaching an adequate understanding of the topic that the question represents.

Fredericksen (1984) charges that, as a result of this process, test results misinform the evaluation of educational systems. Performance in a school

subject may *appear* to improve, because test scores rise, but the apparent improvements may simply result from the targeting of instruction towards tested skills and the growth in knowledge of test demands. The inference from improving test scores, that there has been a proportionate growth in student learning, may not be justified as the improvements may not generalise to areas of the focal construct that go untested. Empirical support for this view is provided by a series of studies in the United States which uncovered what became known as the Lake Wobegon effect (Koretz 1988). Improvements in school pupils' test scores over time, which resulted in all states performing above published national averages, failed to generalise to alternative measures of the same constructs (Cannell 1988, Koretz 1988, Linn, Graue and Sanders 1990).

Linn (2000) describes the effect of introducing a new test into an educational system. At first, scores are comparatively low, but during a period of adjustment in the school system, the scores rise steadily as teachers and learners adapt to the demands of the test. When the first test is replaced with a new, unfamiliar measure, scores fall. They then rise once again as teachers and learners adapt to the demands of the new test.

Score-boosting practices which fail to develop the range of construct skills have been dubbed *test score pollution* (Haladyna et al 1991) because they lead to mistaken inferences regarding ability. Haladyna et al list a variety of supposedly unethical test-preparation practices which may lead to score pollution including, among others, developing a curriculum or teaching objectives based on test items, presenting items similar to those on the test and using score-boosting activities. In their view such practices should be disallowed because they have led to a situation in which test results have come to misrepresent the outcomes of public education.

While many psychometricians may regard these test-preparation practices as questionable (Mehrens and Kaminsky 1989), unethical (Haladyna et al 1991) or immoral (Cannell 1988), this perspective places an unreasonable burden on teachers. Wiliam (1996) reminds us that teachers rightly consider it their duty to obtain the best test results for their students. Teachers and students often believe that tests contain what should be learned and therefore what must be taught: they do not distinguish between the target domain and test content. A view that Chapelle and Douglas (1993) regard as perfectly reasonable, given that a test such as IELTS represents the language hurdle students must clear before pursuing their academic careers.

Following this line of argument, Davies (1985, 1990) suggests that high-stakes proficiency tests, although founded on theoretical constructs and intended to be curriculum-free, will inevitably attract to themselves a syllabus and hence evolve into achievement tests. He considers any attempt to prevent teaching to the test both futile and misguided. Washback is so widespread that it is more rational to accept it and then work to make it as

beneficial as it can be so that its influence on the classroom is transformed (Davies 1990). Based on the many examples of negative effects, it is assumed that similar mechanisms can be exploited to promote good practice when the technology of testing is reformed. This pragmatic notion of tests worth teaching to can be traced back at least as far as Thorndike (1921: 378), who writes, 'Students will work for marks and degrees if we have them. We can have none, or we can have such as are worth working for.'

In this spirit the distinction is often made between intentional, beneficial or positive washback associated with innovative test methods and negative washback engendered by inappropriate or outmoded forms of assessment which fail to keep pace with developments in pedagogy (Davies 1990, Bailey 1996, Hughes 2003).

As Chapman and Snyder (2000) point out, a number of features of test design may be manipulated in efforts to improve instruction. These include *item format* (multiple-choice, short-answer question, extended response etc.), *content* (topics and skills), *level of knowledge called for* (retention, understanding or use), *complexity* (the number of content areas and their interrelationship), *difficulty* (easy or challenging), *discrimination* (in terms of set standards of performance), *referential source* (criterion-referenced or norm-referenced), *purpose* (learner performance, curriculum evaluation, teacher evaluation) and *type of items* (proficiency, achievement or aptitude).

Although precise descriptions of how tests have been reformed to promote washback are often lacking (Cheng 2005, Wall 2005), Hughes (2003) devotes a chapter to achieving beneficial washback and Brown (2000) summarises suggestions for the promotion of positive washback from Hughes (2003), Heyneman and Ransom (1990), Shohamy (1992), Kellaghan and Greaney (1992), Bailey (1996), and Wall (1996). Brown categorises these prescriptions as *test design strategies*, *test content strategies*, *logistical strategies* and *interpretation strategies*.

Test design and content strategies are more closely identified with washback direction, while logistical issues are more closely identified with washback intensity (Kellaghan and Greaney 1992, Hughes 1993). Interpretation strategies may be viewed as indirect, policy-level means of ensuring standards of test design and logistical provision while the test design and content strategies relate most closely to Chapman and Snyder's (2000) test description categories of format, content, complexity and referential source. The following sections consider how these test design features are said to engender washback.

Test format

Madaus (1988) cites evidence of teachers limiting the tasks undertaken in the classroom to the types of task set in a test. He proposes as one of seven

general principles for test impact that teachers will pay close attention to the format of questions on a test (for example, essay or multiple choice) and will adjust their teaching accordingly (Madaus 1988). Much of the criticism in EFL directed at tests such as the TOEFL has concentrated on the use of a multiple-choice item format and norm-referenced score interpretations (Spolsky 1995, Hughes 2003).

Multiple-choice tests have come in for particular criticism for negative washback on the grounds that they may restrict test content, atomise knowledge and encourage poor teaching practices (Wise 1985, Resnick and Resnick 1992, Prodromou 1995, Hughes 2003). However, others have rejected the assertions that the multiple-choice format is only capable of addressing *lower order* skills of recall or recognition and believe that recent test instruments employing multiple-choice items to test *higher order* analytical skills demonstrate their point (Wiliam 1996, Mehrens 1998).

On the other hand, Wiliam (1996) suggests that even where multiple-choice items have been designed to assess higher-order thinking, this is achieved by identifying *particular* higher-order skills. So even though the full breadth of a domain may be addressed, multiple-choice questions break it down into manageable, assessable units. Because the domain as a whole never gets assessed, teachers tend to concentrate on isolated elements. This criticism, which relates in Chapman and Snyder's (2000) terms to complexity as well as format, is of particular relevance to current conceptions of language use as 'the dynamic and interactive negotiation of meaning between two or more individuals in a particular situation' (Bachman and Palmer 1996: 62).

Aside from the constraints imposed by multiple-choice items on the content of a test, a further charge is that they affect teaching methods; that multiple-choice testing leads to multiple-choice teaching (Smith 1991a). Hughes (2003) argues that practice in taking multiple-choice items and test-taking strategies will not provide learners with the most effective means of improving their language ability and Smith (1991a, 1991b) cites examples of multiple-choice teaching among primary school teachers in the USA in response to the Iowa Test of Basic Skills. However, it is not immediately clear whether the preponderance of such practices should be attributed to test format or to other shortcomings in the educational system, for example, poor standards of teacher training (Hamp-Lyons 1998).

It is often asserted, as by Hughes (2003), that direct, constructed response item formats, such as the essay writing required in the IELTS Writing Modules, will yield more positive washback than multiple-choice items. However, direct tests of performance are not immune to narrow test-taking strategies. Linn, Baker and Dunbar (1991), in a discussion of validity issues surrounding performance assessment, see the potential for a narrow strategic approach to composition. They assert that more direct forms of assessment will not necessarily foster classroom activities that are more conducive to