1 Impact, washback, evaluation and related concepts: definitions and examples

Impact is the main topic of this book, so this opening chapter will attempt to clarify the concept and its implications. It will also consider related terms, for example *evaluation, monitoring* and *washback*.

Context for the discussion throughout will be via reference to the role of impact studies in the test development and validation systems of the University of Cambridge English for Speakers of Other Languages (ESOL) Examinations.

This book will refer to two Cambridge ESOL impact study projects in particular.

One is the study of the impact of the International English Language Testing System (IELTS), an examination 'designed to assess the English language ability of people whose first language is not English and who need to study, work or live where English is used as the language of communication' (www.ielts.org). The second is a study of the impact of the *Progetto Lingue 2000* (Year 2000 Languages Project), a Ministry of Education, Italy, state school foreign language education improvement programme.

This should set the scene for Chapter 2, which considers different approaches to the collection and analysis of impact data, and Chapter 3, on the definition of research objectives and questions. Chapter 4 then traces the development of impact study instrumentation, and Chapter 5 the collection, management and analysis of data. In Chapters 6 and 7, some of the main findings of the studies into IELTS and the *Progetto Lingue 2000* impacts are presented, in their own right and as examples of the outcomes that may be expected from research into the foreign language learning and testing aspects of educational impact. Chapter 8 traces research and other developments related to the two studies, considers lessons to be learned, and suggests approaches for the continuing study of educational impact.

But first some key *terms* need to be defined.

Impact in educational research

Impact of process and product

Taking an *educational evaluation* viewpoint, Weiss defines *impact* as 'the net effects of a programme (i.e. the gain in outcomes for program participants minus the gain for an equivalent group of non-participants)'(1998: 331). She then broadens this somewhat narrow definition by adding that 'impact may also refer to program effects for the larger community', and admitting that 'more generally it is a synonym for *outcome'* [all italics mine]. This wider view of the impact construct is reflected in a definition from *developmental education* studies:

> Impacts (also referred to as effects) may be planned or unplanned; positive or negative; achieved immediately or only after some time; and sustainable or unsustainable … Impacts may be observable/measurable during implementation, at project completion, or only some time after the project has ended. Different impacts may be experienced by different stakeholders (Department for International Development (DFID) Glossary of terms 1998).

Note that this definition of impact appears to include a focus on *processes* as well as *outcomes* or product, a distinction often at issue in impact and evaluation studies. Roy defines process and product studies as follows:

> A study of the product is expected to indicate the pay-off value while a study of the process is expected to indicate the intrinsic values of the programme. Both are needed, however, to find the worth of the programme (1998:71).

Weiss defines a process focus more straightforwardly as the study of 'what goes on while a program is in progress', whereas outcome studies measure and describe the 'end results of the program' (1998:334–335).

In the field of education, *impact studies* most commonly focus on the effects of interventions, including both teaching programmes and tests, on the people participating in them in various ways. Given the formative nature of education and learning, such studies seek to measure and analyse both outcomes, for example test results or subsequent performance on the criteria the test is measuring, and processes, for example the learning and teaching approaches and activities of programmes preparing candidates for a test.

The study of the impacts of the IELTS test is, by definition, a form of summative evaluation, concerned with outcomes such as candidate test performances. But a test such as IELTS, used as an English language qualification for academic studies in English-speaking countries and for immigration, training and employment purposes, also has significant potential impact on processes such as the ways candidates learn and prepare for the test

itself and for their English language activities beyond it (there is further discussion on this below). There are thus formative aspects (intended to provide information to improve programmes or tests) as well as summative aspects to impact studies. As for the *Progetto Lingue 2000* (*PL2000*) Impact Study, of a Ministry of Education project for the improvement of language learning in the state sector, there is a focus on developments in areas where the Project is intended to have influence. These include, of course, teaching/learning processes and foreign language performance outcomes. Impact studies of tests, like impact studies of learning programmes, are likely to be process- as well as product- or outcome-oriented.

Impact studies and evaluation studies

Varghese contrasts *impact* studies with *evaluation* studies, the latter, he feels, tending to focus more closely on the immediate objectives of projects rather than their longer-term development.

> … An evaluation of adult literacy programmes may indicate the total number of persons made literate by the programme. An impact study of the programme will focus on the social implications of the outcomes … It will also ask, for example, whether the reading habits of the community improved (1998:49).

Varghese (49–50) reminds us that impacts are changes (or effects) rather than the achievement of project targets, which are often seen as the focus of evaluation studies.

Weiss defines evaluation as the 'systematic assessment of the operation and/or outcomes of a program or policy, compared to explicit or implicit standards, in order to contribute to the improvement of the program or policy' (1998:330). The term 'evaluation' refers to an overall process; an evaluation study is, after all, an exercise to appraise (that is, measure the value of) an educational programme. Impact may well be *one of the areas* of the programme covered by an evaluation. So, evaluation and impact are linked, with evaluation in some cases tending to include impact, in the sense of programme effects which evaluators want to find out about as part of their evaluation. Why? In order to make proposals to adjust these effects to 'contribute to the improvement of the program or policy' (see Weiss above).

But the evaluation study literature (e.g. Agar 1986, Connell et al 1995, Cronbach 1982, MacDonald 1974, Parlett and Hamilton 1972, Weiss 1998) warns us regularly that the nature of evaluation should not be over-simplified as it was following Scriven's 1967 contrast between summative and formative evaluation. 'Evaluations that focus on outcomes' says Cronbach, 'can and should be used formatively' (1982). Parlett and Hamilton go further with their concept of 'illuminative evaluation' (1972). This has a primary concern 'with

description and interpretation rather than measurement and prediction', with how innovation operates, 'how it is influenced by the various school situations in which it is applied; what those directly concerned regard as its advantages and disadvantages; and how students' intellectual tasks and academic experiences are most affected'. Illuminative evaluation 'aims to discover and document what it is like to be participating in the scheme' (see Murphy and Torrance (eds.) 1987:60–61). Levine's concept of evolving curriculum is similarly dynamic, 'where evaluation is an inherent aspect of the curriculum planning process *(evaluation in planning)'* with 'the evaluation process itself a perpetual and self-developmental inquiry process *(evaluation as planning).* The curriculum evaluation process that emerges is flexible, yet methodical, open, yet directive, *and* respectful of the diverse, complex curricular visions, needs and constraints encountered in schools and classrooms' (2002:26).

There would seem to be much to learn from these definitions. The impact studies discussed in this book attempt to combine 'description and interpretation' with 'measurement and prediction'. They seek to investigate the influences of 'the various school situations' in which the IELTS test is prepared for and the principles of the *PL2000* are put into practice. They certainly seek to discover 'what those directly concerned' regard as the 'advantages and disadvantages' of the test and the curriculum reform project, and also how students' 'intellectual tasks and academic experiences are most affected'. We shall also see, throughout this book, that the study of the impact of language tests or programmes, like the evaluation process, tends to be 'perpetual and self-developmental' rather than single and monolithic.

Monitoring

Then there is the term *monitoring,* clearly related to both impact and evaluation, and actually suggested by Italian colleagues participating in the study of the impact of the *PL2000*, as a synonym for impact study. Weiss defines monitoring as '[a]n ongoing assessment of program operations conducted during implementation, usually by sponsors or managers, to assess whether activities are being delivered as planned, are reaching the target populations, and are using resources appropriately' (1998:333). Judging by this definition, there is considerable overlap between monitoring and evaluation, but the fact that monitoring takes place only *during* the implementation of a programme may distinguish it.

A further distinction, suggested by Lynda Taylor (2005, personal communication) sees monitoring as primarily descriptive in function, followed by evaluation, which is, naturally, mainly evaluative in function. As will emerge from this chapter (see Figure 1.5 on page 20) the Cambridge

ESOL model of test development includes, following the establishment of the need for a new or revised test, the stages of:

- design and initial specification
- development through trialling, analysis, evaluation and review
- *monitoring*, mainly through routine descriptive data for analysis, until a decision is made, based on particular monitoring information
- reviewing and *evaluating* the test for possible further revision.

Weiss links monitoring with 'process evaluation', and adds a participant dimension:

> … process evaluation is not very different from what is often called monitoring. One key difference is that monitoring is done primarily on behalf of the funders and other high-level officials to hold the program to account (1998:181).

One teacher/administrator participant in the study of IELTS impact seemed to sense that impact studies may be less top-down and judgemental when she described them as more 'user-friendly' than evaluations or monitoring.

In the case of the *PL2000* Impact Study there was, of course, no question of the impact study 'funders', Cambridge ESOL, holding policy-makers, designers, managers or officials of the *PL2000* to account. Rather, the examinations board was concerned with the two-way impacts (see below) of Cambridge exams on participants in the *Progetto,* and of the *Progetto* on these exams. The *PL2000* Impact Study was carried out by Cambridge ESOL as an interested party, selected alongside other international test providers (see below), to provide external certification for students who had been participating in foreign language courses under the *PL2000*. Cambridge ESOL was not the initiator or leader of the foreign language reform project itself; its role, through the *PL2000* Impact Study which it ran with Ministry approval, was to describe impacts rather than to evaluate the Project.

Insider and outsider roles

The question of the 'evaluator-user and insider-outsider interface' is often at issue in the evaluation literature. O'Dwyer summarises as follows:

> Evaluators may remain distant and report findings in their own way with the expectation that these may be used to improve a program; or, may be actively involved in the program, working hand-in-hand with those in a program, or stakeholders to the program, with a view to specifying the evaluation focus according to the needs of the users. The profile which an external evaluator may adopt, therefore, could be of a complete outsider to a program, or, towards the other end of the spectrum, of a close 'insider' in relationship to the clients (2005).

The differences in evaluator roles described here would appear to apply to impact studies as well as evaluations. With the design, trialling and implementational phases of the study of IELTS impact, which will be described in detail in Chapters 3–6 below, the outsider-insider roles included, at various stages, both external consultant teams and individuals commissioned by Cambridge ESOL, and validation and implementation expertise from within the organisation. Cambridge ESOL is, of course, one of the three partners in the IELTS test, along with the British Council and IDP Education Australia : IELTS Australia. Both these latter partners are fully informed of the impact studies and themselves contribute to research in support of IELTS through the IELTS funded-research programme (see this chapter and Chapters 4, 6 and 8). In the *PL2000* Impact Study, it will be seen, relationships between the impact study team and participants such as the case study school teachers and heads were close, though not quite 'insider'.

Impact and washback in foreign language teaching and testing

In this section of the chapter, the concepts of evaluation, monitoring and impact are investigated within the fields of language teaching and testing, where similarities with and distinctions from the general educational literature will be discovered.

In the language teaching and testing literature, the concept of impact as effects or changes still stands but the term co-occurs frequently with the term 'washback' (or 'backwash') and it is the distinction between the two that is often an issue of debate. In the context of studies of the effects of language programmes or tests on those involved, the concepts of impact and washback/backwash are often considered in terms of their:

- logical location
- definition and scope
- positive and negative implications
- intentionality
- complexity
- direction
- intensity, emphasis
- stakes and stakeholders
- relationships with validity and validation
- relationships with the Critical Language Testing view
- role in impact/washback models.

This chapter attempts below to cover all these aspects of impact and washback.

Washback and impact

'Washback and the impact of tests more generally has become a major area of study within educational research' Alderson (2004a:ix) and as the washback and impact net widens, so does the need for agreed labels for the kinds of study we carry out to investigate the effects of tests or programmes *in and beyond the classroom context*. Hamp-Lyons summarises the situation and the terminology well. She finds that Alderson and Wall's 'limitation of the term 'washback' to influences on teaching, teachers, and learning (including curriculum and materials) seems now to be generally accepted, and the discussion of wider influences of tests is codified under the term 'impact' (Wall 1997), which is the term used in the wider educational measurement literature' (2000:586). In similar vein, Bachman and Palmer 1996 refer to issues of test use and social impact as 'macro' issues of impact, while washback takes place at the 'micro' level of participants, mainly learners and teachers.

So the term 'impact' now appears to be used to describe studies which investigate the influences of language programmes and/or tests on stakeholders *beyond* the immediate learning programme context. An impact study might, for example, investigate the effects of a programme or test on school heads, parents, receiving institution administrators, high-stakes test providers (all these stakeholders included in the two impact studies described in Chapters 3–8 below).

Given that the term 'impact' is a word in everyday use in its meaning of 'influence or effect' (e.g. *Oxford School Dictionary,* 1994), it is unsurprising to find the term also apparently used non-technically. When Alderson (2004a: ix), for example, writes: 'We now know, for instance, that tests will have more impact on the content of teaching and the materials that are used than they will on the teacher's methodology', is he using the term in its lay sense, since technically the content of teaching and the teacher's methodology are washback rather than impact matters? Or is he acknowledging that, for some, impact, the broader construct, *includes* washback? Green notes that although 'the terms have been used to refer to the same concept, backwash is distinguished from test impact by Bachman and Palmer (1996:30) who, with McNamara (1996, 2000), Hamp-Lyons (1998) and Shohamy (2001) place washback within the scope of impact' (2003:6). This would presumably mean that one could use the term 'impact' for all cases of influence from a language test or language programme, whether on teaching and learning or on, say, a university's admissions policy.

Andrews, writing on washback and curriculum innovation, appears to acknowledge the fragility of the washback : impact distinction:

> The term washback is interpreted broadly … the present chapter uses *washback* to refer to the effects of tests on teaching and learning, the educational system, and the various stakeholders in the education process. Where the word 'impact' occurs in this chapter, it is used in a non-technical sense, as a synonym for 'effect' (2004:37).

In this book we shall try to be consistent in the use of terms:

- to use 'washback' to cover influences of language tests or programmes on language learners and teachers, language learning and teaching processes (including materials) and outcomes
- to use 'impact' to cover influences of language tests or programmes on stakeholders beyond language learners, teachers, except when it is the influences of a test or programme on learners and teachers *outside* their learning or teaching roles, for example on their attitudes to matters beyond language learning; in this case the book will tend to refer to impact e.g. Research Question 4: What is the impact of IELTS on the participants who have taken the test?

In terms of these definitions, the two studies which are the focus of this book cover both washback and impact. They are called 'impact studies' because of this breadth.

Washback/backwash

Hamp-Lyons notes that washback 'is one of a set of terms that have been used in general education, language education and language testing to refer to a set of beliefs about the relationship between testing and teaching and learning' (1997:295). Another of the 'set of terms' is 'backwash', but it would appear that the terms 'washback' and 'backwash' are used interchangeably in the field. '… to clarify the distinction between the terms backwash and washback', Alderson says (2004a:xi), 'there is none'. Hughes admits that there is interchangeable use of the two terms in his work but adds, (2003:57) 'Where "washback" came from I do not know. What I do know is that I can find "backwash" in dictionaries, but not "washback"'. Cheng and Curtis choose to use the term 'washback' 'as it is the most commonly used in the field of applied linguistics' (2004:5). This book will follow suit, preferring the term 'washback' as it does now appear to be in more common use in the field.

Impact and validity

Saville and Hawkey cite 'the implementation of new national curricula with national achievement tests' (2004:75) in the UK and New Zealand as examples of a growing tendency for tests to be used to provide evidence of and targets for change, thus having more significant influences on the lives of individuals and groups.' Language tests such as IELTS are more and more frequently used in a 'gate-keeping' role in decisions of crucial importance to candidates such as the admission or otherwise to particular programmes, professions or places. They thus earn the label of 'high-stakes' tests. The social consequences of test use are a growing concern.

Messick insists on the inclusion of the outside influences or 'consequential validity' of a test in its validation, 'the function or outcome of the testing, being either interpretation or use' (1989:20). In an interesting personal communication to Alderson, however, Messick warns against too glib a view of the relationship between test washback or impact and test validation.

> Washback is a consequence of testing that bears on validity only if it can be evidentially shown to be an effect of the test and not of other forces operative on the educational scene … Washback is not simply good or bad teaching or learning practice that might occur with or without the test, but rather good or bad practice that is evidentially linked to the introduction of the use of the test (Alderson 1995:3).

Alderson (1995:4) himself takes 'an agnostic position' on the relationship between test impact and test validity. He agrees that 'test consequences are important and may relate to validity issues (bias being perhaps the most obvious case)' but has 'difficulty seeing the washback and impact as central to construct validity' because of the 'myriad factors' impacting on a test: teacher's linguistic ability, training, motivation, course hours, class size, extra lessons and so on. 'This is not, of course, to deny,' Alderson notes in his paper written for Phase One of the study of IELTS impact (see also Chapters 2 and 4), 'the value of studying test impact and washback in its own right, but it underscores the need to gather evidence for the relationship between a test and its impact on the one hand, and of the futility, given current understanding and data, of making direct and simplistic links between washback and validity'Alderson (1995:3).

Green agrees that backwash 'is not generally considered to be a standard for judging the validity of a test', because 'backwash can only be related to a test indirectly, as effects are realised through the interactions between, *inter alia,* the test, teachers and learners' (2003a). Green cites Mehrens (1998) on 'the lack of agreed standards for evaluating backwash' and the fact that 'different stakeholders may regard the same effects differently'. There is interesting food for thought in Messick's 1996 advice, also cited by Green: 'rather than seeking backwash as a sign of test validity, seek validity by design as a likely basis for backwash' (1996:252).

Hamp-Lyons links the increasing importance attached to tests to the washback/impact relationship, claiming that the 'shift from washback to impact suggests a growing awareness by language testers that the societies in which and for which we work are, whether we approve or not, using tests as their levers for social and educational reform' (2000:586). Actually, as Alan Davies points out (personal communication), this is by no means a new phenomenon, being a feature, for example, of the Civil Service examinations in India in the 19th century.

But Bachman (2004) still feels that validity and test use are not necessarily accepted as related in language assessment, despite Messick (1989) and Bachman's own earlier view (1990). In the Bachman and Palmer 1996 definition of test 'usefulness', entailing six qualities: reliability, construct validity, authenticity, interactiveness, *impact* and practicality, 'both the construct validity of our score-based inferences and the impact, or consequences, of test use need to be considered from the very beginning of test design, with the test developer and test users working together to prioritise the relative importance of these qualities' (2004:5). Bachman considers that considerations of validity and impact are thus subsumed 'under a unitary concept of test usefulness'. In Chapter 8, Bachman's case for the articulation of assessment use arguments, in terms of claims, warrants, backing and rebuttals is discussed. These could well feature in a model of test impact study.

Whether impact is intended or unintended, it would seem to be a legitimate and crucial focus of research, both micro and macro, to 'review and change' tests and programmes in the light of findings on, among other aspects of programmes or tests, 'how the stakeholders use the exams and what they think about them' (Saville 2003:60). This is a justification, of course, for studies of the effects of exams as part of the test *validation process*, that is 'the process of investigating the quality of test-based inferences, often in order to improve this basis and hence the quality of the test' (McNamara 2000:138).

The location of impact studies in programme and test development

Figure 1.1 suggests a sequence of washback and impact events in the context of a new educational programme.