Does Washback Exist?

J. CHARLES ALDERSON and DIANNE WALL

Lancaster University

The notion of 'washback' is common in the language teaching and testing literature, and tests are held to be powerful determiners of what happens in classrooms. Claims are made for both negative and positive washback, and some writers go so far as to claim that a test's validity should be measured by the degree to which it has a beneficial effect on teaching. However, very little evidence has been presented to support the argument that tests influence teaching, and what evidence has appeared tends to be based on teachers' accounts of what happens in the classroom rather than on observations of teaching and learning. This article explores the notion of washback and advances a series of possible Washback Hypotheses. It then reviews the empirical research in general education and in language education to see what insights can be gained into whether washback actually exists, how it can be measured, and what accounts for the form it takes. The article concludes with a series of proposals for further research into a phenomenon on whose importance all seem to be agreed, but whose nature and presence have been little studied.

INTRODUCTION

The notion that testing influences teaching is commonplace in the educational and applied linguistics literature. This phenomenon is referred to as 'backwash' in general education circles, but it has come to be known as 'washback' in British applied linguistics. We will use the term 'washback', as this is the tradition in our field, but we see no reason, semantic or pragmatic, for preferring either label.

Many educationalists have written about the power of examinations over what takes place in the classroom (see, for example, Vernon 1956; Davies 1968; Kellaghan, Madaus, and Airasian 1982; Alderson 1986; Morrow 1986; Pearson 1988; Hughes 1989; Khaniya 1990a and 1990b, among others). Pearson, for example, says: 'It is generally accepted that public examinations influence the attitudes, behaviour, and motivation of teachers, learners, and parents' (1988: 98). This influence is often seen as negative: Vernon (1956: 166) claimed that examinations 'distort the curriculum'. He felt that teachers tended to ignore subjects and activities which did not contribute directly to passing the exam, and lamented what he considered to be excessive coaching for exams. Others, however, see washback in a more positive way. Morris (1972: 75) considers examinations necessary to ensure that the curriculum is put into effect. Swain (1985: 42–4) recommends that test developers 'bias for best' and 'work for washback', while Alderson (1986: 104) argues for innovations in the language curriculum through innovations in language testing.

Applied Linguistics, Vol. 14, No. 2 © Oxford University Press 1993

Pearson takes this idea further and claims not only that good tests will encourage the use of 'beneficial teaching–learning processes', but also that they

> will be more or less directly usable as teaching–learning activities. Similarly, good teaching–learning tasks will be more or less directly usable for testing purposes, even though practical or finncial constraints limit the possibilities. (Pearson 1988: 107)

Some writers have even gone so far as to suggest that a test's validity should be measured by the degree to which it has had a beneficial influence on teaching. Morrow (1986) coined the term 'washback validity' to denote the quality of the relationship between a test and associated teaching. The notion presumably means something like: 'this test is valid when it has good washback'; and conversely, 'this test is invalid when it has negative washback'. He says: 'The first validity criterion that I would . . . put forward for [these examinations] would be a measure of how far the intended washback effect was actually being met in practice' (Morrow 1986: 6). He admits, however: 'I am not sure at this stage how it could be measured', although he then goes on to claim: 'In essence an examination of washback validity would take testing researchers into the classroom in order to observe the effect of their tests in action.' He cites Wilkins, Widdowson, and others as asserting that direct tests of language performance will be 'most beneficial in terms of washback effect', and argues that communicative tests like the former Royal Society of Arts Examination in the Communicative Use of English as a Foreign Language should have a 'powerful and positive washback effect into the classroom'.

Frederiksen and Collins (1989) introduce a concept similar to 'washback validity'. They use the term 'systemic validity', which they define as follows:

> A systemically valid test is one that induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure. Evidence for systemic validity would be an improvement in those skills after the test has been in place within the educational system for a period of time (1989: 27)

However, to our knowledge, this form of validity has never been demonstrated, or indeed investigated, nor have proposals been made as to how it could be established empirically rather than asserted. Moreover, it is not at all clear that if a test does not have the desired washback this is necessarily due to a lack of validity of the test, as Morrow and Frederiksen and Collins simplistically imply. It is surely conceivable that other forces exist within society, education, and schools that might prevent washback from appearing, or that might affect the nature of washback despite the 'communicative' quality of a test. This can then hardly be attributed to a problem with the test. Whereas validity is a property of a test, in relation to its use, we argue that washback, if it exists—which has yet to be established—is likely to be a complex phenomenon which cannot be related directly to a test's validity.

It seems to us to be important to investigate the nature of washback first, and the conditions under which it operates. Only once we are able to describe what

actually happens, will we be in a position to explore what 'causes' these effects. And only after we have established causal relationships will we be in a position to explore whether we are justified in relating washback to a test's validity. Thus, talk of washback or systemic validity is at best premature, and at worst ill-conceived.

In summary, the term 'washback' is common in the language teaching and testing literature and tests are held to be powerful determiners of what happens in classrooms. However, the concept is not well defined, and we believe that it is important to be more precise about what washback might be before we can investigate its nature and whether it is a natural or inevitable consequence of testing.

This article is in several sections: first, we speculate upon some possible interpretations of washback. Then, we refer to the general educational literature for enlightenment, since the applied linguistics literature appears to take washback for granted rather than to question it. Next, we look at what empirical research exists in the language testing field for the insights it can offer. Finally, we briefly present a series of proposals for further research.

EXPLORING THE CONCEPT OF WASHBACK

The term 'washback' is itself a neutral one, and can be related to 'influence'. If the test is 'poor', then the washback may be felt to be negative. But if the Washback Hypothesis holds, then good tests should have good effects (as yet undefined) rather than negative effects.

If we consider these beliefs briefly, we can see that other possibilities also hold. The Washback Hypothesis seems to assume that teachers and learners do things *they would not necessarily otherwise do* because of the test. Hence the notion of influence. But this also implies that a 'poor' test could conceivably have a 'good' effect if it made teachers and learners do 'good' things they would not otherwise do: for example, prepare lessons more thoroughly, do their homework, take the subject being tested more seriously, and so on. And indeed, teachers are often said to use tests to get their students to do things they would not otherwise do: to pay attention to the lesson, to prepare more thoroughly, to learn by heart, and so on. To the extent that these activities are in some sense desirable—hard work is presumably more 'desirable' than no work at all, and extrinsic motivation might be better than no motivation at all—then any test, good or bad, can be said to be having beneficial washback if it increases such activity or motivation.

Alternatively, one might wish to consider the possibility of a test, good or bad, having negative effects. The most obvious such effect is anxiety in the learner brought about by having to take a test of whatever nature, and, if not anxiety, then at least concern in teachers, if they believe that some consequence will follow on poor performance by the pupils. The argument would go like this: any learner who is obliged to do something under pressure will perform abnormally and may therefore experience anxiety. Thus pressure produces abnormal performance, the fear of which produces anxiety. In addition, the fear of the

consequences of particular performances produces anxiety which will influence those performances. Similarly for teachers, the fear of poor results, and the associated guilt, shame, or embarrassment, might lead to the desire for their pupils to achieve high scores in whatever way seems possible. This might lead to 'teaching to the test', with an undesirable 'narrowing of the curriculum' (Smith 1991).

We may also wish to consider the possibility of a test *reinforcing* some behaviour or attitude rather than bringing about an otherwise unlikely behaviour. Thus students may already work hard, and a test may simply motivate them to work harder. A learner may constantly self-evaluate against internal or external criteria, and the test may provide very useful additional criteria for this kind of comparison. Thus the relationship between a test and its impact, positive or negative, might be less simple than at first sight appears to be the case. The quality of the washback might be independent of the quality of the test.

The question arises as to whether 'washback' is the same as 'influence' or whether the term refers solely to some sorts of influence and not others? We might not want to use the term 'washback' for the anxiety caused by having to take a test, but might well want to apply it to syllabus or textbook design specifically based on a test (for example, the Longman series of textbooks intended to prepare students for the Cambridge First Certificate in English examination—see, for example, Kingsbury 1983).

Even if we were to use the term 'washback' to refer to the test's effect on textbook design, we would probably need to distinguish between pedagogic material which is directly related to a specific test in terms of content or method (see, for example, the Kennedy, Kenyon, and Matthiesen 1989 preparation course for the Test of English as a Foreign Language (TOEFL)) and material which is intended to help students get ready for an exam in some more general way—for example, study skills courses which claim they give students skills relevant to taking a test of English for Academic Purposes like the International English Language Testing System (IELTS). Given these complexities, we may wish to restrict the use of the term 'washback' to classroom behaviours of teachers and learners rather than to the nature of printed and other pedagogic material. It is not clear from the literature, however, that writers do indeed make these distinctions.

Another aspect of the notion of washback that needs to be explored is its deterministic nature. How directly, according to the Washback Hypothesis, do tests bring about change in teaching and learning? A naïve deterministic view (which is often implicit in the complaints about TOEFL, for example, or even in the claim that tests can be used as 'levers for change') would assume that the fact of a test having a set of qualities is sufficient in itself, by virtue of the importance of tests in most societies, to bring about change. However, this takes little account of other factors in the environment which may also influence teaching: the teachers' basic competence, their understanding of the principles underlying the test, levels of resourcing within the school system, etc.

Most discussions of washback tend to assume that the existence of a test brings about some change in motivation and thus in behaviour. In fact, the relationship between motivation and performance is a very complex matter, beyond the scope of this discussion. However, a thorough study of washback must surely take account of research findings in this area. In fact, there appear to be conflicting results, as Fransson's brief review (1984) indicates. Fransson points out that an increase in level of motivation is accompanied by an increase in learning, up to an optimal point. However, beyond that point an increase in motivation seems to have negative effects and performance declines (the so-called 'Yerkes-Dodson Law'). The position of the optimal point, Fransson suggests, depends upon the difficulty of the task. However, it may well also relate to the consequences of the task (in our case, the test), as well as to other factors within the performer such as that person's need for achievement (nAch) (McDonough 1981: 146). This may be the result of two opposed tendencies: the motivation to succeed and the motivation to avoid failure. These, in their turn, may relate to the person's expectations of success (or failure), the value of the task as an incentive, and the person's orientation toward success or toward avoidance of failure.

As if this were not sufficiently complicated, McDonough goes on to review a further theoretical position, that of attribution theory, which describes motivated behaviour in terms of 'the causes to which the individuals attribute or ascribe their own and other people's performance: their own ability, effort, intention or others' ability, effort and intention, luck and so on' (ibid.: 147).

It may be, however, that the key factor is not motivation but anxiety, both 'state anxiety'—the condition associated with performing a task—and 'trait anxiety'—one's habitual response to stress. Furthermore, it may be important to distinguish two sorts of anxiety: debilitating and facilitating. Which of these is aroused in a particular learner or teacher may depend on personality factors (for example, extroversion/introversion, need for achievement, fear of failure, and so on) as well as the consequences (and the learners' perception of those consequences) of particular performances.

What this brief excursion into motivation and anxiety is intended to illustrate is the extreme complexity of the topic, and the contrasting naïvety of the Washback Hypothesis: clearly, those asserting the existence of washback need to take more account of research findings and resulting theoretical positions in related fields.

The point we are making is that the Washback Hypothesis is unduly simplistic and makes too many untested assumptions about how people are influenced. This applies as much to negative washback as it does to positive washback. However, it will be important when empirically examining washback to look at both negative and positive situations, to see how comparable they are.

THE WASHBACK HYPOTHESIS

It might help to clarify our thinking if we attempt to state the Washback Hypothesis explicitly. From a reading of the literature on language testing generally,

and from our experience of talking to teachers about their teaching and testing, it is possible to develop different hypotheses, from the most general and vague to the somewhat more refined, which take account of different factors. It is useful to try to separate out the factors, as below.

Some possible Washback Hypotheses

(1) *A test will influence teaching.* This is the Washback Hypothesis at its most general. However, a second partly different hypothesis follows by implication from this first one, on the assumption that teaching and learning are related, but not identical:

(2) *A test will influence learning.*

Since it is possible, at least in principle, to separate the content of teaching from its methodology, then we need to distinguish the influence of a test on the content of the teaching from its influence on the methodology. Thus:

(3) *A test will influence* ***what*** *teachers teach*; and

(4) *A test will influence* ***how*** *teachers teach*; and therefore by extension from (2) above:

(5) *A test will influence* ***what*** *learners learn*; and

(6) *A test will influence* ***how*** *learners learn.*

However, perhaps we need to be somewhat more precise about teaching and learning, in order to consider how quickly and in what order teachers teach and learners learn. Hence:

(7) *A test will influence the* ***rate*** *and* ***sequence*** *of teaching*; and

(8) *A test will influence the* ***rate*** *and* ***sequence*** *of learning.*

Similarly, we may wish to consider explicitly both the quality and the quantity of teaching and learning:

(9) *A test will influence the* ***degree*** *and* ***depth*** *of teaching*; and

(10) *A test will influence the* ***degree*** *and* ***depth*** *of learning.*

If washback relates to attitudes as well as to behaviours, then:

(11) *A test will influence attitudes to the content, method, etc. of teaching and learning.*

In the above, however, no consideration has been given to the nature of the test, or to the uses to which scores will be put. Yet it seems not unreasonable to hypothesize:

(12) *Tests that have important consequences will have washback*; and conversely

(13) *Tests that do not have important consequences will have no washback.*

It may be the case that:

(14) *Tests will have washback on **all** learners and teachers.*

However, given what we know about differences among people, it is surely likely that:

(15) *Tests will have washback effects for **some** learners and **some** teachers, but **not** for others.*

Clearly, we are complicating what was initially a simple assumption. Is this justified? Is washback a concept to be taken seriously, or simply a metaphor which is useful in that it encourages us to explore the role of tests in learning and the relationship between teaching and testing? We are not sure at present, but we suspect that if it is a metaphor, it needs to be articulated somewhat more precisely if it is to throw light on teaching and testing, or indeed on the nature of innovation and change. And if it is a concept to be taken seriously, then we need to examine it critically, and see what evidence there might be that could help us in this examination.

We need in either case to identify and examine cases where washback is thought to have occurred, and to see how and why it did or did not occur.

RESEARCH INTO WASHBACK

The general educational literature

Surprisingly little empirical research has been conducted into the nature or, indeed, the existence of washback in general education, much less in language education. Even much-cited authorities such as Vernon (1956) tend to make general statements about washback rather than referring to evidence from specific studies, and even where studies are quoted—for example, with details of the effect of coaching on examination performance—no references are given which would enable an independent evaluation of that evidence or a description of the nature of the coaching.

The 'classic' study in the general education literature is that by Kellaghan, Madaus, and Airasian, into 'The Effects of Standardized Testing', published in 1982. This joint Irish–American investigation examined the impact on Irish schools of introducing standardized tests.

Ireland was chosen because, unlike the USA and elsewhere, there was no tradition of standardized testing in existence. It was thus possible to introduce such tests selectively into experimental schools and contrast outcomes and attitudes with control groups of various sorts.

The study took place from 1974 to 1977. In the experimental group, standardized tests were given to pupils in Grades 2 to 6, and resulting norm-referenced information was given to teachers. In one control group, no testing was carried out. In the other, the tests were administered, but no results were given to teachers. The study investigated school-, teacher-, pupil-, and parent-level effects. At school level, the study looked at the effect of tests on school

organization and practice, and on school achievement. At teacher level, it looked at teachers' attitudes to standardized tests, and their reported use of test information. The researchers looked at pupils' reactions to standardized tests, their perceptions of factors that affect scholastic progress, their self-concepts, and their self-assessments. They also investigated parents' familiarity with changes in evaluation, the communication between school and parents, parents' perception of their children's school progress, and their knowledge about and attitudes toward standardized testing. In addition, the study investigated the role of test information in teacher expectations and their perceptions of pupils.

The results showed very little effect of standardized tests on school organizational or assessment practices—they tended to be used to support rather than to disrupt existing practices. Nothing in the findings supported the belief that taking tests regularly leads to increased performance on later tests: a comparison of experimental and control groups' performances revealed complex results that could be interpreted in a variety of ways, but which did not provide evidence of a simple cause–effect relationship (Kellaghan *et al.* 1982: 61).

The reactions of teachers and pupils to the test programme were very positive. The tests were perceived as fair and accurate, and teachers who had received tests and test information were more positive about tests and their value than were their colleagues in the control groups. Pupils took the tests seriously and a large majority reported enjoyment of the experience rather than fear. There was no evidence that the provision of test information had negative effects on their self-concepts, or their self-assessments. Parents showed very little impact of the experimental tests: they were largely unaware of the existence of such tests, although their attitudes to testing in general were very positive. There was evidence of an expectancy effect, such that teachers in receipt of test information about their pupils rated their pupils in line with that test information. However, expectancy effects were at work regardless of whether teachers had test information or other information about their pupils.

These findings seemed to suggest that the introduction of standardized tests was beneficial to the schools, or at least not harmful. However, one criticism is that the situation was artificial: the tests were introduced for the sake of the experiment and had no currency or consequence within the Irish educational system. Pupils were not passed or failed, or denied entry to further education on the basis of their test results. It is therefore not surprising that relatively little negative impact was perceived. The Washback Hypothesis presumably applies to tests and examinations that are used regularly within the curriculum and which are perceived to have educational consequences. To such settings, the Kellaghan *et al.* study has little of direct relevance.

A second drawback of the Kellaghan *et al.* study from the point of view of this paper is that the dependent variables were either teacher ratings of pupils, test scores, or questionnaire responses. Very little usable information was gathered independently on what happened in the experimental and control-group classrooms, and we are obliged to rely upon teacher and pupil accounts of practices. Although the study is valuable for what it reveals with respect to a variety of

aspects of test impact, particularly in relation to perceptions, observational evidence of test impact on classroom teaching/learning is minimal.

Indeed, this lack of evidence from classrooms is a characteristic of virtually all writings about the influence of tests on teaching. See, for example, Paris, Lawton, Turner, and Roth (1991); Haladyna, Nolen, and Haas (1991); or Frederiksen (1984)—all of whom use anecdote, assertion, or interviews and surveys of what teachers and pupils say they do rather than direct observation.

An exception is Smith (1991), who reports on two qualitative studies which investigated the effect of tests on teachers and classrooms. Data from interviews revealed that the publication of test results induced feelings of fear, guilt, shame, embarrassment, and anger in teachers, and the determination to do what was necessary to avoid such feelings in the future. Teachers believed that test scores were used against them, despite the perceived invalidity of the scores, and they also believed that testing had severe emotional impact on young children (less so on older pupils). From classroom observation, it was concluded that testing programmes substantially reduced the time available for instruction and narrowed the curriculum and modes of instruction:

> What we saw in one school's sixth grade was a transition, as the school year progressed toward ITBS testing in April, from laboratory, hands-on instruction in science several days a week, to less frequent science out of textbooks (choral reading from the text and answering comprehension and vocabulary questions on worksheets), to no science instruction at all in the weeks before the test, to either no science at all or science for entertainment value during the ITBS recovery phase, to science instruction precisely tailored to the questions in the district criterion-referenced tests, to no science at all. The same group devoted about 40 minutes each day to writing projects in the fall, but the class wrote no more after January, after which they spent the time on worksheets covering grammar, capitalization, punctuation, and usage. Writing instruction returned in late May, when the pupils again began producing poetry, stories, reports on projects for the short time remaining in the school year. Social studies and health instruction disappeared altogether. (Smith 1991: 10)

Interestingly, however, Smith reports that there were two different reactions to this 'narrowing of the curriculum'. One was accommodation by teachers, who discarded what was not going to be tested, and taught towards the test, but the other was one of resistance, exemplified by one teacher:

> I know what's on the test, but I feel that these children should keep up with current events and trace the history behind what's happening now, so we're going to spend March doing that. I guess I'm saying that the test scores are going to be up for grabs. (ibid.)

This suggests that the washback phenomenon is not quite as simple as is at times made out. We need many more studies like those Smith reports before we can claim we understand the nature and mechanisms of washback.

Language education

There is remarkably little research in the field of language education that can be said to have investigated and established what washback is or how it works.

Much assertion exists—for example, the debate about the influence of the introduction of multiple-choice tests in Ethiopia: see Forbes (1973) and Madsen's reply (Madsen 1976). Forbes attacks objective test methods, and makes claims about what happens in Ethiopian classrooms:

> Gone are the happy days in which a teacher could spend a whole period on his [*sic*] favourite poem, 'The Solitary Reaper', if he wanted to. He may not even spend time on Belloc's 'Tarantella', even though it is in the prescribed textbook written by some of the university 'English language specialists' ... So it's eyes up to the sentences on the blackboard—sentence patterns for tenses, for quantifiers, for modals, for relative clauses. Which is right and which is wrong? Write them down to remember them, perhaps, but don't write anything else. That's waste of time [*sic*]. We are back to 'The pen of my aunt is in the garden' and 'The postillion was struck by lightning' with a vengeance. (Forbes 1973: 135)

Sadly (but perhaps not surprisingly), Forbes provides no evidence to support his emotive claims, nor does Madsen in his reply. Even when justifying the introduction of objective tests in terms of how teachers were preparing students for the old examinations that the objective tests replaced, Madsen has to resort to impressions:

> Teachers *appeared* to be short-changing their students in the classrooms. English teachers in the upper grades in particular *seemed* to be spending virtually all their time on examination techniques rather than on the English fundamentals so badly needed. (Madsen 1976: 136, our italics)

Similarly, when describing the effect of the objective test, unsupported claims are made:

> Teachers in the upper grades *were inclined to* model instruction on the now sacrosanct objective examination ... the backwash effect on the schools became just as devastating as that produced by the earlier précis-essay examination. (Madsen 1976: 138, our italics)

The only projects that are known to the current authors in language education that have systematically investigated the phenomenon relate to the Netherlands (Wesdorp 1982), Turkey (Hughes 1988), Nepal (Khaniya 1990a, 1990b), and Sri Lanka (Alderson, Clapham, and Wall 1987; Pearson 1988; Alderson and Wall 1990, 1991; Wall 1991).

Information, published or unpublished, may exist with respect to other countries. It is nevertheless noteworthy that we have failed to uncover more empirical studies, given the firmness with which a belief in washback is held in language teaching circles.

The Netherlands. Wesdorp (1982) gives an unpublished account of research into the validity of objections to the introduction of multiple-choice tests into the assessment of mother tongue and foreign language education. The research found that most of the objections, which assumed washback effects, were not justified. It was, for example, assumed that skills that could not be tested by

multiple-choice (mcq) would not be taught in primary schools, but a comparison of essays written before the introduction of mcq writing tests, and twelve years after that introduction, found no differences in quality. An investigation of differences in teacher activities in schools with and without an mcq final test failed to show any clear differences. No evidence was forthcoming of an increased use of mcq in language teaching, nor of any change in student study habits as a result of mcq tests in English. In short, empirical investigation revealed much less washback effect than had been feared.

Turkey. Hughes (1988) describes a project at Boğaziçi University, Istanbul, where innovations were made in test design with a view to bringing about change in the curriculum. Prior to the start of the project, students were entering mainstream academic studies (taught in the medium of English) with very low levels of English proficiency, despite a year's preparation at the Foreign Language School (FLS). Because of this poor English performance, it was decided to design a new proficiency test, which would have to be passed by students who wish to transfer from the FLS to the University. The content of the test was meant to reflect the sorts of uses of English that might be expected in an English-medium university like Boğaziçi. The immediate reported effect was that:

> for the first time, at least for some years, FLS teachers were compelled, by the test, to consider seriously just how to provide their students with training appropriate for the tasks that would face them at the end of the course. (Hughes ibid.: 144)

The evidence was that standards of English did indeed rise, since by the end of the first year in which the new proficiency test was introduced, 83 per cent of students achieved the minimum acceptable Michigan score (compared with fewer than 50 per cent before the new test), and a survey of academic staff showed that the English proficiency of students entering mainstream studies was perceived to be 'very, very, much better' than their predecessors. Hughes claims that this state of affairs came about because of the beneficial washback effect of the test: 'Teaching for the test (which may be regarded as inevitable) became teaching towards the proper objectives of the course' (ibid.: 145), since the test was based directly on the English language needs of undergraduate students. Hughes seems to demonstrate that tests can indeed impact on the language curriculum, especially if their consequences are important, as in the case of the Bogazici proficiency test. Changes in the syllabus, textbooks, and possibly in the teaching in the FLS are reported to have occurred. It also seems to be the case that something associated with these changes brought about improved levels of proficiency in English—although what that something was, is unclear. It is at least conceivable that *the mere threat* that students might actually be failed on a proficiency test of whatever nature led to students and teachers working harder, but not necessarily in the 'right' (i.e. intended) direction.

Despite the fact that the new proficiency test was quite unlike the Michigan test in content and method, preparation of students for the new test resulted in

increased proficiency on a very different test. Hughes unfortunately does not discuss this issue. Nor do we know what actually changed in classrooms; although increases in English proficiency were established, the origin of these is uncertain. In short, we do not know what washback effect the test produced, nor how it produced it.

Nepal. In Nepal the SLC (School Leaving Certificate) is an extremely important hurdle to tertiary education, good employment, and social status. Khaniya (1990b) describes the existence of published cribs for the exam with exotic titles like 'Gautam Super Lucky SLC Guess Paper' and 'Guess Paper with High Surety for SLC'. The SLC as described requires students to memorize texts and answers to questions, since many of the test questions and texts are taken directly from the textbooks, and are actually not answerable without reference to the textbook (or a memorized version of it). In such circumstances, where memorization is essential to successful performance, it is perhaps not at all surprising that exam coaching occurs, and visible signs of this, like the publications mentioned, are clear evidence of washback in some form. However, we have no description of how teachers actually teach to the exam, what and how students learn, and so on. In fact, Khaniya reports very high failure rates on the SLC (90 per cent), which must mean that if cramming is necessary for the exam, and if, as he asserts, cramming is rife, it must either be very inefficient, or the exam must be more unpredictable than the writers of cribs and cramming courses admit, or than Khaniya himself describes. What we clearly need, and lack, is a description of what this claimed washback actually looks like and how and when it is successful.

Khaniya's attempt at investigating washback is, in fact, indirect. He designed a new exam, on 'communicative' lines, intending to engineer beneficial washback, and then compared his new exam with the traditional SLC. He gave his exam to students at the beginning and at the end of Grade 10 (when students are preparing for the SLC). Since they did not on the average improve in their performance on the new exam, Khaniya claims that this is because students are cramming for the SLC, which cramming does not 'teach English'. Unfortunately, Khaniya was not able to administer the SLC at the beginning and end of the year as well: it could be that the ability to take SLC did not improve either.

Sri Lanka. Studies of the classroom impact of examinations are very rare indeed. The only study identified to date is the Sri Lankan O-Level Evaluation Project (Wall and Alderson 1993). This project investigated the effects on Sri Lankan classes of changing the O-Level English examination. These changes were intended to reinforce innovations in textbook materials and teacher training courses. Wall and Alderson conducted a two-year longitudinal observational study of O-Level English classes, together with studies of teacher and pupil attitudes to the new examination.

Their conclusion is that the impact of the new examination is less pervasive than had been expected. Although the examination has had a demonstrable

effect on the content of language lessons, it has had virtually no impact on the way that teachers teach English. The longitudinal observations revealed no changes in teaching methodology over the two years of the study, which were compared with baseline data gathered before the new examination was introduced. Wall and Alderson suggest this lack of impact may be because of a lack of understanding on the part of teachers of what might be an appropriate way to prepare students for the examination. It may even be that the exam itself—and this may indeed be true of all exams—does not and cannot *determine how* teachers teach, however much it might influence *what* they teach. This has important consequences for the nature of the Washback Hypothesis.

A SERIES OF PROPOSALS FOR RESEARCH

Clearly, more research is needed in this area. We have already suggested that it is important to define what is meant by the term 'washback': what scope it should have, where its limits lie, and what aspects of impact we might not wish to include in the concept of washback. Secondly, it is important to state explicitly what one's version of the Washback Hypothesis is: it is highly likely that it will be more complex than the fifteen hypotheses put forward above. It will be necessary to spell out in some detail what the predicted effects of the test are, and it is highly likely that this statement will have to take account of the nature of the test concerned, the educational context within which it is used, and the nature of the decisions that are taken on the basis of the test results.

In addition, in parallel to this increasing specification of the Washback Hypothesis, it will be important to take account of findings in the research literature in at least two areas: that of motivation and performance, and that of innovation and change in educational settings.

Then we will need to consider the methodology to be used in research into washback. There has been a tendency to date to rely upon participants' reported perceptions of events through interview and questionnaire responses, or to examine the results and relationships of test performances. Smith is instructive with respect to possible methods: 'We employed direct observation of classrooms, meetings and school life generally; interviews with teachers, pupils, administrators, and others; and analysis of documents' (1991: 8). We suggest that it is increasingly obvious that we need to look closely at classroom events in particular, in order to see whether what teachers and learners say they do is reflected in their behaviour.

In addition, we believe it important in conjunction with classroom observations to triangulate the researcher's perceptions of events with some account from participants of how they perceived and reacted to events in class, as well as outside class—this amounts to an advocacy of a more ethnographic approach to the topic than has been common to date (see Watson-Gegeo 1988 for a discussion of such an approach).

Finally, as well as attempting to *describe* the washback that occurs, we need to attempt, at some point in the future, to *account for* what occurs, and this is likely to involve widening our hypothesis formulation and data collection to include

explanatory variables derived from the research literature in motivation and innovation.

What this amounts to is a long-term and relatively complex research programme. We believe that this is both inevitable and desirable. What is undesirable is a continuation of our state of ignorance about a phenomenon on whose importance all seem to be agreed. Equally undesirable is a continuation of naïve assertions about washback on the part of applied linguists in general and materials writers, syllabus designers, and language teachers in particular, as well as language testers, until some empirical investigations have been undertaken.

(Revised version received October 1992)

REFERENCES

Alderson, J. Charles. 1986. 'Innovations in Language Testing?' in Portal (ed.), pp. 93–105.

Alderson, J. Charles, C. Clapham, and **D. Wall.** 1987. *An Evaluation of the National Certificate in English*. Institute for English Language Education, Lancaster University.

Alderson, J. Charles and **D. Wall.** 1990. *The Sri Lankan O-Level Evaluation Project: Second Interim Report*. Lancaster University.

Alderson, J. Charles and **D. Wall.** 1991. *The Sri Lankan O-Level Evaluation Project: Third Interim Report*. Lancaster University.

Chamberlain, D. and **R. Baumgardner** (eds.). 1988. *ESP in the Classroom: Practice and Evaluation*. (ELT Document 128) Modern English Publications.

Davies, A. 1968. *Language Testing Symposium: A Psycholinguistic Approach*. Oxford: Oxford University Press.

Forbes, D. 1973. 'Selling English Short.' *English Language Teaching Journal* 27: 132–7.

Fransson, A. 1984. 'Cramming or understanding? Effects of intrinsic and extrinsic motivation on approach to learning and test performance' in J. Charles Alderson and A. H. Urquhart (eds.): *Reading in a Foreign Language*. London: Longman.

Frederiksen, N. 1984. 'The Real Test Bias.' *American Psychologist* March pp. 193–202.

Frederiksen, J. R. and **A. Collins.** 1989. 'A systems approach to educational testing.' *Educational Researcher* 18/9: 27–32.

Haladyna, T. H., S. B. Nolen, and **N. S. Haas.** 1991. 'Raising standardized achievement test scores and the origins of test score pollution.' *Educational Researcher* 20/5: 2–7.

Hughes, A. 1988. 'Introducing a needs-based test of English language proficiency into an English-medium university in Turkey' in A. Hughes (ed.).

Hughes, A. (ed.). 1988. *Testing English for University Study*. (ELT Document 127) Modern English Publications.

Hughes, A. 1989. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Kellaghan, T., G. F. Madaus, and **P. W. Airasian.** 1982. *The Effects of Standardized Testing*. London: Kluwen, Nijholf Publishing.

Kennedy, D. B., D. M. Kenyon, and **S. J. Matthiesen.** 1989. *Preparing for the TOEFL*. Cambridge, Mass.: Newbury House.

Khaniya, T. R. 1990a. 'The Washback Effect of a textbook-based test.' *Edinburgh Working Papers in Applied Linguistics* 1: 48–58.

Khaniya, T. R. 1990b. 'Examinations as Instruments for Educational Change: Investigating the Washback Effect of the Nepalese English Exams.' PhD thesis, University of Edinburgh.

Kingsbury, R. 1983. *Longman First Certificate Coursebook*. London: Longman.

Lee, Y. P., C. Y. Y. Fok, R. Lord, and **G. Low** (eds.). 1985. *New Directions in Language Testing*. Hong Kong: Pergamon Press.

McDonough, S. 1981. *Psychology in Foreign Language Teaching*. Hemel Hempstead: Allen and Unwin.

Madsen, H. 1976. 'New alternatives in EFL exams or "How to avoid selling English short".' *English Language Teaching Journal* 30/2: 135–44.

Morris, B. 1972. *Objectives and Perspectives in Education: Studies in Educational Theories*. London: Routledge and Kegan Paul.

Morrow, K. 1986. 'The evaluation of tests of communicative performance' in Portal (ed.).

Paris, S. G., T. A. Lawton, J. C. Turner, and **J. L. Roth.** 1991. 'A developmental perspective on standardized achievement testing.' *Educational Researcher* 20/5: 12–19.

Pearson, I. 1988. 'Tests as levers for change' in Chamberlain and Baumgardner (eds.).

Portal, M. (ed.). 1986. *Innovations in Language Testing*. London: NFER/Nelson.

Smith, M. L. 1991. 'Put to the test: the effects of external testing on teachers.' *Educational Researcher* 20/5: 8–11.

Swain, M. 1985. 'Large-scale communicative testing' in Lee *et al.* (eds.).

Vernon, P. E. 1956. *The Measurement of Abilities*. (2nd. edn.) London: University of London Press.

Wall, D. 1991. 'Measuring Examination "Washback": The Sri Lankan Evaluation Project.' Paper presented at the IAEA Conference, Nairobi, Kenya, May.

Wall, D. and **J. Charles Alderson** 1993. 'Examining Washback: The Sri Lankan Impact Study.' *Language Testing* 10/1.

Watson-Gegeo, K. A. 1988. 'Ethnography in ESL: defining the essentials.' *TESOL Quarterly* 22/4: 575–92.

Wesdorp, H. 1982. *Backwash effects of language testing in primary and secondary education*. Stichting Centrum voor onderwijsonderzoek van de Universiteit van Amsterdam.