TOEFL®

Monograph Series

MS - 15
JUNE 1999

Washback in Language Testing

Kathleen M. Bailey

ETS Educational Testing Service

Washback in Language Testing

Kathleen M. Bailey

Educational Testing Service
Princeton, New Jersey
RM-99-4

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

To obtain more information about TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org

Web site: http://www.toefl.org

Foreword

The TOEFL® Monograph Series features commissioned papers and reports for TOEFL 2000 and other Test of English as a Foreign Language program development efforts. As part of the foundation for the TOEFL 2000 project, a number of papers and reports were commissioned from experts within the fields of measurement and language teaching and testing. The resulting critical reviews and expert opinions were invited to inform TOEFL program development efforts with respect to test construct, test user needs, and test delivery. Opinions expressed in these papers are those of the authors and do not necessarily reflect the views or intentions of the TOEFL program.

These monographs are also of general scholarly interest, and the TOEFL program is pleased to make them available to colleagues in the fields of language teaching and testing and international student admissions in higher education.

The TOEFL 2000 project is a broad effort under which language testing at ETS will evolve into the 21st century. As a first step in the evolution of TOEFL language testing, the TOEFL program recently revised the Test of Spoken English (TSE®) and announced plans to introduce a TOEFL computer-based test (TOEFL CBT) in 1998. The revised TSE test, introduced in July 1995, is based on an underlying construct of communicative language ability and represents a process approach to test validation. The TOEFL CBT will take advantage of the new forms of assessments and improved services made possible by computer-based testing while also moving the program toward its longer-range goals, which include

- the development of a conceptual framework that takes into account models of communicative competence
- a research agenda that informs and supports this emerging framework
- a better understanding of the kinds of information test users need and want from the TOEFL test
- a better understanding of the technological capabilities for delivery of TOEFL tests into the next century

It is expected that the TOEFL 2000 efforts will continue to produce a set of improved language tests that recognize the dynamic, evolutionary nature of assessment practices and that promote responsiveness to test user needs. As future papers and projects are completed, monographs will continue to be released to the public in this new TOEFL research publication series.

TOEFL Program Office
Educational Testing Service

Abstract

This monograph summarizes recent research on language testing washback. It begins by compiling several definitions of washback and related constructs. It then poses a model of language testing washback and examines the available research related to this model. The monograph concludes with recommendations for appropriate research methods to be used in future investigations of washback.

Acknowledgments

In writing this review of the literature, I have been influenced by the work of a large number of people, but most especially Charles Alderson, Arthur Hughes, Elana Shohamy, and Dianne Wall.

I would also like to acknowledge the help of Liying Cheng, Liz Hamp-Lyons, and Carol Taylor, each of whom helped me find hard-to-locate research on language testing washback.

Table of Contents

Figure

Biographical Note

Kathleen M. Bailey is a professor of Applied Linguistics at the Monterey Institute of International Studies in Monterey, California. She received her MA in Teaching English as a Second Language and her Ph.D. in Applied Linguistics from the University of California at Los Angeles.

Dr. Bailey spent the 1996-97 academic year at the English Language Teaching Unit of the Chinese University of Hong Kong, where much of this work was completed. Her research interests include language assessment, teacher development, second language acquisition, and classroom research.

Part I: Introduction

It has long been asserted, in a wide range of contexts, that tests exert a powerful influence on language learners who are preparing to take these exams, and on the teachers who try to help them prepare. (See Spolsky, 1994 for a historical overview.) The following statements are typical of claims in the literature:

> It is generally accepted that public examinations influence the attitudes, behavior, and motivation of teachers, learners and parents . . . (Pearson, 1988, p. 98)
>
> It is common to claim the existence of washback (the impact of a test on teaching) and to declare that tests can be powerful determiners, both positively and negatively, of what happens in classrooms. (Wall & Alderson, 1993, p. 41)
>
> The washback effects of large-scale testing programs on instruction are widely discussed. In the view of instructors and students, such tests contain what students must learn and therefore what must be taught – a reasonable view, given that the tests in many cases represent the language hurdle students must clear before continuing their academic careers. (Chapelle & Douglas, 1993, p. 16)

Swain (1985) succinctly states the prevailing opinion: "It has frequently been noted that teachers will teach to a test: that is, if they know the content of a test and/or the format of a test, they will teach their students accordingly" (p. 43).

Tests are often perceived as exerting a conservative force which impedes progress. As Andrews and Fullilove (1994) point out, "Not only have many tests failed to change, but they have continued to exert a powerful negative washback effect on teaching" (p. 57). These authors also note that "educationalists often decry the 'negative' washback effects of examinations and regard washback as an impediment to educational reform or 'progressive' innovation in schools" (ibid., pp. 59-60). As Heyneman (1987) has commented, "It's true that teachers teach to an examination. National officials have three choices with regard to this 'backwash effect': they can fight it, ignore it, or use it" (p. 260; as cited in Andrews, 1994a, p. 51).

The problem is that while washback is widely perceived to exit, there is little data to confirm or deny these perceptions. This is neatly summarized by Alderson and Hamp-Lyons (1996) in the rationale for their study of TOEFL preparation classes in the United States: "Much has been written about the influence of testing on teaching. To date, however, little empirical evidence is available to support the assertions of either positive or negative washback" (p. 281).

Andrews (1994a) concurs: "Although a great deal has been said and written about washback, there is in fact relatively little empirical evidence for its existence" (p. 44).

Similarly, Shohamy (1993a) acknowledges that "while the connection between testing and learning is commonly made, it is not known whether it really exists and, if it does, what the nature of its effect is" (p. 4).

In 1993, Alderson and Wall noted that the existence and importance of washback had been widely asserted, but that relatively few empirical studies had been done to investigate the nature of the construct. However, three years later, in the introduction to a special issue of *Language Testing* that was devoted to the topic of washback, they were able to write,

> Concern has long been voiced about the power of tests to affect what goes on in the classroom, the educational system, and society as a whole – the so-called "washback effect." However, it was not until recently that language testers began to take a critical look at the notion of washback and to try to determine whether it did in fact exist, whether it was predictable, what form it might take, and what explanations there were for its appearance or absence in particular settings. The last five years have seen a growing interest in the phenomenon, both on the theoretical and empirical fronts . . . (1996, p. 239)

It is this "growing interest" in washback that provides the basis for this literature review.

In this monograph I will explore and summarize recent research on second language testing washback, first by compiling definitions of washback and related constructs. (For reviews of the research on washback in first language testing environments, see Andrews, 1994a; Cheng, 1997, forthcoming; Fredericksen, 1984; and Herman & Golan, 1993.) Then I will use a model suggested by Hughes (1993) and printed in Bailey (1996a) as a springboard for discussing the participants, the processes, and the products involved in the washback effect. Finally, I will comment on potential research methods for investigating phenomena associated with washback, and suggest research that could be done, using the TOEFL 2000 as an example of a large-scale, high-stakes test revision project – the implementation of which provides an ideal opportunity for investigating possible washback.

Part II: Defining and Describing Washback

In recent years an explosion of research on washback has led to a greater understanding of this construct than was previously available. In this section I will first review definitions of *washback* and related concepts. Then I will discuss the washback construct as a criterion for evaluating and developing language tests. Finally, I will present a model of washback which provides a useful framework for the review of empirical research in the following sections of this monograph.

Definitions of Washback and Related Concepts

Definitions of washback are nearly as numerous as the people who write about it. These definitions range from simple and straight-forward to very complex. Some take a narrow focus on teachers and learners in classroom settings, while others include reference to tests' influences on educational systems and even on society in general. Some descriptions stress intentionality while others refer to the apparently haphazard and often unpredictable nature of washback.

In an important paper on testing listening comprehension, Buck (1988) describes the apparent effect of Japanese university entrance examinations on English-language learning in Japan. He describes washback as follows:

> There is a natural tendency for both teachers and students to tailor their classroom activities to the demands of the test, especially when the test is very important to the future of the students, and pass rates are used as a measure of teacher success. This influence of the test on the classroom (referred to as *washback* by language testers) is, of course, very important; this washback effect can be either beneficial or harmful. (p. 17)

Thus Buck's definition stresses the impact of a test on what teachers and students do in classrooms. (See also Fullilove, 1992, p. 131)

Shohamy (1992) also focuses on washback in terms of language learners as test-takers when she describes "the utilization of external language tests to affect and drive foreign language learning in the school context" (p. 513). She notes that "this phenomenon is the result of the strong authority of external testing and the major impact it has *on the lives of test takers*" (ibid., emphasis added). Shohamy cites as examples the introduction of new English speaking tests in Israel (see Shohamy, Reves, & Bejarano, 1986) and of the *ACTFL Guidelines and Oral Proficiency Interview* in the United States. She states that these two examples both involve "the power of tests to change the behavior of teachers and students" (Shohamy, 1992, p. 514).

In a later paper, Shohamy (1993a, p. 4) summarized four key definitions that are useful in understanding the washback concept:

1. *Washback effect* refers to the impact that tests have on teaching and learning.

2. *Measurement driven instruction* refers to the notion that tests should drive learning.

3. *Curriculum alignment* focuses on the connection between testing and the teaching syllabus.

4. *Systemic validity* implies the integration of tests into the educational system and the need to demonstrate that the introduction of a new test can improve learning.

These ideas are revisited in Shohamy, Donitsa-Schmidt, and Ferman (1996, p. 298), and reviewed by Cheng (1997, p. 39), who also introduces the concept of *washback intensity*: "the degree of washback effect in an area or a number of areas of teaching and learning affected by an examination" (ibid., p. 43).

In another article, Shohamy (1993b) contrasts school tests and external tests. She notes that

> external tests have become most powerful devices, capable of changing and prescribing the behaviour of those affected by their results – administrators, teachers and students. Central agencies and decision makers, aware of the authoritative power of external tests, have often used them to impose new curricula, textbooks and teaching methods. Thus external tests are currently used to motivate students to study, teachers to teach, and principals to modify the curriculum. The use of external tests as a device for affecting the educational process is often referred to as the *washback effect* or *measurement-driven instruction.* (p. 186)

Shohamy goes on to describe a "collaborative/diagnostic feedback model" in which school personnel and test developers work together to purposefully create a feedback loop in which testing influences teaching and vice versa.

Berry (1994) also notes an increased interest in washback with her definition: "One of the major issues within the field of assessment in the 1990s has been a concern with the systemic validity of tests – the so-called 'washback effect' or the effect a test has on classroom practice" (p. 31). Thus although Berry's definition takes a narrow focus on the classroom, she combines the notion of washback with systemic validity, where other writers (including Shohamy, 1993a) have made a distinction between these two terms.

Pierce takes a broader view, but one that is somewhat similar to Berry's. Pierce (1992) states that "the washback effect, sometimes referred to as the systemic validity of a test, . . . refers to the impact of a test on classroom pedagogy, curriculum development, and educational policy" (p. 687).

Cohen (1994) also takes a broad view. He describes washback in terms of "how assessment instruments affect educational practices and beliefs" (p. 41). We will see that these broad categories cover the two main research foci for investigations of washback: actions and perceptions.

More recently, Bachman and Palmer (1996, pp. 29-35) have discussed washback as a subset of a test's impact on society, educational systems, and individuals. They state that test impact operates at two levels: the micro level (i.e., the effect of the test on individual students and teachers) and the macro level (the impact on society and its educational systems). The following comment from Buck (1988) illustrates how these two levels often work in tandem:

> Japan is a country in which the entrance examination reigns supreme. It is almost impossible to overstate the influence of these examinations on both the educational system as a whole, and the day-to-day content of classroom teaching. Their importance in the lives of young people is such that almost all future social and economic advancement is dependent on the results of these entrance examinations. (p. 16)

Bachman and Palmer (1996, p. 35) note, however, that washback is a more complex phenomenon than simply the effect of a test on teaching and learning. Instead, they feel the impact of a test should be evaluated with reference to the contextual variables of society's goals and values, the educational system in which the test is used, and the potential outcomes of its use.

Andrews (1994a) sees washback as "a complex and ill-defined phenomenon" (p. 45). He adds another dimension to the definition in terms of the scope of people influenced by test results, when he acknowledges "widespread acceptance of the assertion that tests, especially public examinations, exert an influence on teachers, learners, *and parents*, with an associated impact on what happens in classrooms" (ibid.; emphasis added). Andrews stresses the need for more research into washback, in order to better understand and manage the presumed influence of tests, particularly in the realm of promoting curricular innovation.

Hughes (1989, p. 1) states simply that "the effect of testing on teaching and learning is known as *backwash*" (and this term, as he uses it, is synonymous to *washback*, which I will use in this monograph). Hughes' textbook on testing for language teachers includes a brief chapter about promoting beneficial backwash, in which he lists the following suggestions (ibid., pp. 44-47):

1. Test the abilities whose development you want to encourage.
2. Sample widely and unpredictably.
3. Use direct testing.
4. Make testing criterion-referenced.
5. Base achievement on objectives.
6. Ensure [that the] test is known and understood by students and teachers.
7. Where necessary, provide assistance to teachers.

This advice is based on Hughes' own research (see, e.g., Hughes, 1988), as well as his experience as both a test developer and teacher educator.

Messick (1996) makes the more specific point that 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 and use of the test" (p. 254; emphasis added). He points out that

tests which promote positive washback are likely to include tasks which are *criterion samples* – that is, "authentic and direct samples of the communicative behaviors of listening, speaking, reading and writing of the language being learnt" (ibid., p. 241), and he adds that the transition from learning exercises to test exercises "should be seamless" (ibid.). Messick pinpoints one of the main conceptual and methodological difficulties in studying washback: "It is problematic to claim evidence of test washback if a logical or evidential link cannot be forged between the teaching or learning outcomes and the test properties thought to influence them" (ibid., p. 247; see also Messick, 1994).

Alderson and Wall (1993) pose 15 possible hypotheses that they hope will lead to the eventual refinement of the washback construct in empirical investigations (pp. 120-121):

1. A test will influence teaching.
2. A test will influence learning.
3. A test will influence what teachers teach; and
4. A test will influence how teachers teach; and by extension from (2) above,
5. A test will influence what learners learn; and
6. A test will influence how learners learn.
7. A test will influence the rate and sequence of teaching; and
8. A test will influence the rate and sequence of learning.
9. A test will influence the degree and depth of teaching; and
10. A test will influence the degree and depth of learning.
11. A test will influence attitudes to the content, method, etc. of teaching and learning.
12. Tests that have important consequences will have washback; and conversely,
13. Tests that do not have important consequences will have no washback.
14. Tests will have washback on all learners and teachers.
15. Tests will have washback effects for some learners and some teachers, but not for others.

Alderson and Wall posed these hypotheses as a result of their own extensive work in Sri Lanka and of reviewing case studies conducted in Nepal (Khaniya, 1990), Turkey (Hughes, 1988), and the Netherlands (Wesdorp, 1982).

Lam (1994) used Alderson and Wall's (1993) specifications of the washback hypothesis to investigate particular types of washback presumed to have been generated by the Revised Use of English (RUE) exam in the Hong Kong secondary schools. The RUE is "a public examination used principally to evaluate language competence for entrance to Hong Kong's tertiary institutions" (Lam, 1994, p. 84). Lam identified seven different possible outcomes as a result of the change in this major high-stakes examination: *timetable washback, methodology washback, attitude washback, proofreading washback, textbook washback, content washback*, and *performance washback* (ibid., pp. 84-85).

Cheng (1997) also conducted research in the context of secondary school exams in Hong Kong, investigating the Hong Kong Certificate of Education Examination (HKCEE). She uses the term *washback* to mean "an active direction and function of intended curriculum change by means of the change of public examinations" (ibid., p. 38). However, while the revision of the HKCEE syllabus was meant to generate "top-down intended washback on English language teaching and learning in Hong Kong secondary schools" (ibid.), Cheng also acknowledges that "unintended and accidental side-effects could occur" (ibid., p. 39).

Prior to its implementation, Smallwood (1994) expressed concern about the revised HKCEE. While Cheng focused on the exam's intended positive washback, Smallwood predicted that "the proposed changes to the examination may not be the best use of the very powerful washback effect and still may not be the best way to evaluate either the learners' abilities or the appropriacy of the syllabus" (p. 68).

Smallwood argued that continuous assessment was by far preferable to a single, high-stakes examination. He felt that assessment should be based in the classroom, should be conducted in familiar surroundings, and should be ongoing. The washback concept provided the rationale for his position: "this approach is likely to have a real effect on the actual teaching styles used in the classroom regarding the encouragement of oral production by the students in a wide variety of contexts" – a classroom focus that, he acknowledges, was not typically encouraged and was sometimes even discouraged in Hong Kong (ibid., p. 70). Thus Smallwood raised serious concerns about the potential of the revised HKCEE to do harm (i.e., to promote negative washback) in the very area in which it was intended to bring about improvement (positive washback). In another paper about the examination system in Hong Kong, Fullilove (1992) noted that "some critics of the system argue that Hong Kong presents a case of the examination tail wagging the education dog" (p. 31).

Pearson (1988) wrote about the intentional use of a revised national exam in Sri Lanka (see also Wall, 1996; Wall & Alderson, 1993) to bring about curricular change:

> There is an explicit intention to use tests, including public examinations, as levers which will persuade teachers and learners to pay serious attention to communicative skills and to teaching-learning activities that are more likely to be helpful in the development of such skills. (p. 106)

Pearson notes, however, that using "tests as a deliberate backwash-generating device has its limitations" (ibid.).

Smallwood's and Pearson's concerns have the value of refocusing our attention on the intended outcomes versus the incidental outcomes of revising a widely used exam or instituting a totally new exam. The intentionality built into Cheng's definition contrasts with the haphazard nature of the construct implied in a traditional dictionary definition of *backwash*, which Spolsky (1994) quotes as "an unexpected and usually undesirable, subsidiary result or reaction" (p. 55). Spolsky continues: "Strictly speaking, then, the term is better applied only to accidental side-effects of examinations, and not to those effects intended when the *first* purpose of the examination is control of the curriculum" (ibid.). Spolsky states that he can find no support in the dictionary for use of the term *washback*, even though it has been popularized in language testing.

Likewise Andrews (1994a) comments on the *backwash* versus *washback* nomenclature. Based on his own review of the literature, he comments that "in general education literature, the favoured term to describe this phenomenon is 'backwash,' while in language education there seems to be a preference for 'washback'"(p. 67).

We can see from these definitions that while washback is widely held to exist, there are some discrepancies in how it has been defined (and even in what it is called). Nevertheless, its importance in test construction and evaluation should not be underestimated. We turn now to a review of the literature that discusses language testing washback as an evaluative criterion.

Washback as a Criterion for Developing and Evaluating Language Tests

Positive washback, by whatever name, has recently been recognized as one of the main criteria for evaluating language tests. In his 1979 book, *Language Tests at School*, Oller identifies the key characteristics of a good test as being reliability, validity, practicality (also called "feasibility"), and instructional value – the last being most closely related to current conceptions of washback.

In 1988 Hughes wrote about the introduction of a needs-based test of English at the university level in Turkey. This article reports on a fascinating case study of test development and implementation in the face of serious resistance. (Specific findings will be discussed below.) Hughes concluded the paper by saying "potential backwash effect should join validity and reliability in the balance against practicality" (ibid., p. 146).

In discussing the development and implementation of a new national exam in Hong Kong, which would include a direct test of speaking ability, Andrews and Fullilove (1994) expressed the test development team's concern that "as far as possible the test should embody the characteristics of a 'good' test. In particular, [the test development group] kept in mind considerations of validity (especially face and content validity), reliability and washback" (p. 64). Indeed, these authors state that the decision to include a costly oral component in this high-stakes exam represented a desire to enhance its validity and improve the positive washback effect it was expected to exert.

Washback has received even more attention as an evaluative criterion recently, with the advent of communicative language testing. For instance, one of Morrow's (1991) five criteria for evaluating communicative language tests is the idea that such tests should "reflect and encourage good classroom

practice" (p. 111). In describing a test development project called the *Communicative Use of English as a Foreign Language*, Morrow states: "This [i.e., washback] is a major concern underlying the design of tests; indeed in many ways the tests themselves have drawn on 'good' classroom practice in an attempt to disseminate this to other classrooms" (ibid.). Morrow says that a "conscious feedback loop between teaching and testing, in terms of not only content but of approach, is a vital mechanism for educational development" (ibid.; see also Shohamy, 1993b).

The washback concept in language test development has been actively utilized by test developers working in Canada. One influential use of washback in communicative language testing comes from the test development team at the Ontario Institute for Studies in Education, which created the secondary school French test *A Vous La Parole*. (See Canale & Swain, 1980; Green, 1985; Hart, Lapkin, & Swain, 1987; Swain, 1984, 1985, for more information.) One of the team's four tenets for communicative test design was the principle, "Work for washback" – the notion that communicative tests should be explicitly designed to bring about positive washback (Green, 1985, pp. 218-223). Wesche (1987, cited in Pierce, 1992) reports that promoting positive washback was also a major concern in the development of the *Ontario Test of English as a Second Language.*

Boyle and Falvey (1994) observe that "there has been a recent renewal of interest in the link between good teaching and good testing" (p. xi). They also note that washback, along with validity, reliability and practicality, is now "one of the Big Four considerations in evaluating the worth of a test" (ibid.).

From this review of the literature, we can see that language testing washback (1) has often been discussed; (2) is widely held to exist; (3) that there are differing points of view about what the construct may encompass; and (4) that positive washback is viewed as an important criterion in the development and evaluation of language tests. However, until recently very little empirical research has investigated the phenomenon in detail.

A Model of Washback

The following parts of this monograph review the available empirical research on washback in language testing. To provide a structure for this review, I will use a model of washback that is based on a framework suggested by Hughes (1993): "In order to clarify our thinking about backwash, it is helpful, I believe, to distinguish between participants, process and product in teaching and learning, recognizing that all three may be affected by the nature of a test" (p. 2).

In the Hughes framework, *participants* include language learners and teachers, administrators, materials developers, and publishers, "all of whose perceptions and attitudes toward their work may be affected by a test" (ibid.). The term *process* covers "any actions taken by the participants which may contribute to the process of learning" (ibid.). According to Hughes, such processes include materials development, syllabus design, changes in teaching methods or content, learning and/or test-taking strategies, etc. Finally, in Hughes' framework, *product* refers to "what is learned (facts, skills, etc.) and the quality of learning (fluency, etc.)" (ibid.). He continues,

> The trichotomy into participants, process and product allows us to construct a basic model of backwash. The nature of a test may first affect the perceptions and attitudes of the participants towards their teaching and learning tasks. These perceptions and attitudes in turn may affect what the participants do in carrying out their work (process), including practicing the kind of items that are to be found in the test, which will affect the learning outcomes, the product of that work. (ibid.)

Here Hughes stresses the participants' perceptions and attitudes and how these factors affect what they do.

Hughes' suggestion that we consider the participants, processes, and products in examining washback led me to draft the following model (Bailey, 1996a, p. 264):

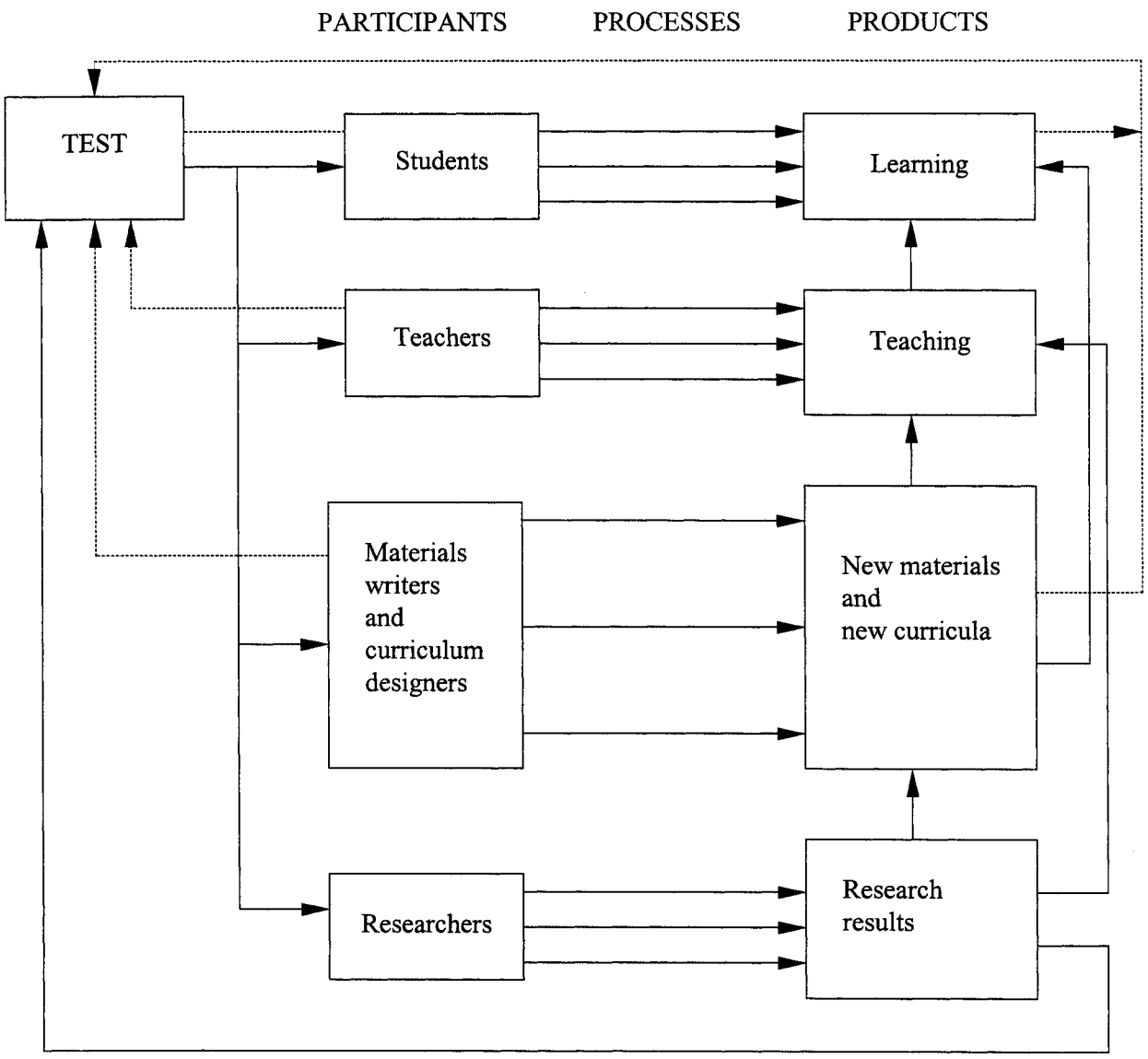

Figure 1: A basic model of washback (reprinted from Bailey, 1996a, p. 264)

Of course, the ultimate product of beneficial washback is, as Hughes states, the improved learning of the construct being measured (language proficiency in our case). But not all of the participants' processes lead directly to learning. Indeed, it is a truism in these days of learner-centered instruction to say that only the learners can do the learning (see, e.g., Nunan, 1996a). The other participants' processes yield ancillary products that we hope will contribute to and promote students' learning – products such as new materials and curricula, improved teaching, valuable research findings, etc.

In the following two sections, I will examine each of Hughes' three components in a review of the available empirical research on washback in language testing. I will situate the findings of that literature within the parameters suggested by this model. The area that has been most closely studied to date involves the washback effect on participants — the focus of the next section.

Part III: Research on Participants in Washback Process

Some kinds of washback result from the effects of a test on the language learners themselves, while other kinds of washback are more closely related to effects of a test on personnel involved in language teaching (including influences on teachers, administrators, course designers, and materials developers – ultimately influencing courses, programs and materials). As noted by Shohamy, Donitsa-Schmidt, and Ferman (1996): "Results obtained from tests can have serious consequences for individuals as well as programmes, since many crucial decisions are made on the basis of test results" (p. 299). I have called these two sorts of washback "learner washback" and "program washback," respectively (Bailey, 1996a). This idea overlaps, to some extent, Bachman and Palmer's (1996, pp. 30-31) micro and macro levels of washback, although they included the influences on individual teachers under the micro category.

The language learners as well as the other participants affected by washback may be influenced by official information about a test prior to its administration (including advertising materials from the test publisher, existing test preparation booklets, etc.), or by folk-knowledge (such as reports from students who have taken earlier versions of the test). They may also be influenced by several sources of feedback following the administration of the test. These would include (but not be limited to) the actual test scores provided by the exam scoring service, feedback from the test-takers (what was easy or difficult, what seemed fair or unfair, unexpected item types, unfamiliar instructions, etc.), feedback from the proctors if the test was administered locally, and feedback from the teachers in reaction to the students' scores. The information might be officially supplied (via score reports and information bulletins), inferred, or even imagined.[1]

Test-takers and Washback

It is worthwhile to sort out the students from the other stakeholders since the washback processes that influence them will directly affect language learning (or non-learning), while the influences on other stakeholders will affect efforts to promote language learning. According to Bachman and Palmer (1996), the test-takers themselves can be affected by (1) "the experience of taking and, in some cases, of preparing for the test; (2) the feedback they receive about their performance on the test; and (3) the decisions that may be made about them on the basis of the test" (p. 31).

Fullilove (1992, p. 138) provides a powerful account of students' experiences taking public examinations in Hong Kong. After describing several "seemingly inexplicable procedures" involved in the exam administration (e.g., "students are allowed to take calculators into English language examinations, but must stow them under their desks"), Fullilove notes that "students often feel that they are very small components of an enormous examination system which is highly impersonal on the one hand but personally highly important on the other" (ibid.).

---

[1] When I was the director of the Intensive English Program at the Monterey Institute of International Studies, it was not unusual for students to tell me that they preferred to take the locally administered Institutional TOEFL because it was "easier" than the International TOEFL. I suspect they were reacting to the familiar surroundings and personnel rather than to actual test length, item types, or content, since the various forms of the locally administered Institutional TOEFL are equated with the International TOEFL for length, content, and item facility. This illustration reminds us of Hughes' emphasis on perceptions and attitudes being key parts of the washback phenomenon.

Five of Alderson and Wall's (1993, pp. 120-121) restatements of the washback hypothesis directly address learner washback:

2. A test will influence learning.

5. A test will influence what learners learn.

6. A test will influence how learners learn.

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

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

Three other parts of the hypothesis refer to both teaching and learning (ibid.):

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

14. Tests will have washback on all learners and teachers.

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

Much more research is needed, however, to see whether and how these washback effects play out in the attitudes and behavior of language learners.

In an earlier article (Bailey, 1996a, pp. 264-265), I suggested that students faced with an important test might participate in (but were not limited to) the following processes:

1. Practicing items similar in format to those on the test.

2. Studying vocabulary and grammar rules.

3. Participating in interactive language practice (e.g., target language conversations).

4. Reading widely in the target language.

5. Listening to noninteractive language (radio, television, practice tapes, etc.).

6. Applying test-taking strategies.

7. Enrolling in test-preparation courses.

8. Requesting guidance in their studying and feedback on their performance.

9. Requesting or demanding unscheduled tutorials or test-preparation classes (in addition to or in lieu of other language classes).

10. Skipping language classes to study for the test.

In my experience as a teacher and program administrator, all of the above have happened regularly.

Although language learners are the key participants whose lives are most directly influenced by language testing washback, there is relatively little research that documents their point of view or their washback-related behavior before and after tests. Some researchers (see, e.g., Cohen, 1984) have reported on what students say about actually taking tests, but more information is needed about learner washback.

Bachman and Palmer (1996) suggest that learners should be included in all phases of test development: "One way to promote the potential for positive impact is through involving test-takers in the design and development of the test, as well as collecting information from them about their perceptions of the test and test tasks" (p. 33). These authors feel that if test-takers are involved in this way, they will perceive tests as more interactive and authentic, and will therefore be more motivated, which could lead to enhanced preparation and hence to better performance.

In some societies, learner washback has important financial implications for pupils and their families, in terms of their access to educational opportunities. For example, Wall and Alderson examined a context in which a new national test was implemented, this time the O-level exams administered at the end of the 11th year of education in Sri Lanka. These authors report that "a student's O-level grades, particularly in English, are among the most important in his or her academic career" (1993, p. 42). In discussing the importance of textbooks in preparing students for national exams in Sri Lanka, Wall and Alderson (ibid., p. 61) describe the time-consuming but widespread practice of having students copy test-type texts from the chalkboard, since books are not available or are too costly. They note that "students from poorer families and in schools with fewer resources were not always able to engage in certain types of exam practice because it took too much time to copy texts from one place to another" (ibid., p. 61).

Ingulsrud (1994) has also commented on the financial impact on students and their families in discussing Japanese university entrance examinations:

> For students who are serious about entering a highly ranked university, a considerable amount of coaching is normal in preparing for the entrance examination. High-school students spend evenings, weekends, and even vacations preparing for the test at the various *juku* [exam preparation schools] that provide a range of coaching services. Supplemental education of this kind costs a good deal of money, and yet students and their families are willing to make such sacrifices. If they do well . . . they are assured of a place in a prestigious university, which, in turn, leads to a successful career in business or government. (pp. 79-80)

Ingulsrud also notes that "it goes without saying that the test preparation industry reflects economic inequalities in education: high-quality coaching is available only to those who can afford it" (ibid., p. 72).

As mentioned above, washback may affect learners' actions and/or their perceptions, and such perceptions may have wide ranging consequences. Sturman used a combination of qualitative and quantitative data to investigate students' reactions to registration and placement procedures at two English-language schools in Japan. The placement procedures included a written test and an interview. He found that the students' perceptions of the accuracy of the placement process (i.e., the face validity of the results) were statistically associated with their later satisfaction with the school, the teachers, and the lessons (1996, p. 347).

Shohamy, Donitsa-Schmidt, and Ferman (1996) report on the stability of the washback effect over time as they investigated two national exams that had been implemented in Israel in the late 1980's. One was a high-stakes test of English as a foreign language (EFL) and the other was a low-stakes examination of Arabic as a second language (ASL). Following Madaus (1990), Shohamy et al. (1996) defines a high-stakes test as one used in a context in which decisions about "admission, promotion, placement or graduation are directly dependent on test scores" (p. 300), while low-stakes exams do not entail these significant decisions. The notion of high-stakes and low-stakes exams is reflected in Alderson's and Wall's (1993) breakdown of the washback hypothesis: "Tests that have important consequences will have washback; and conversely, tests that do not have important consequences will have no washback" (pp. 120-121).

The research conducted in Israel by Shohamy et al. (1996) on the washback effect over time included questionnaires administered to students who were preparing to take either the low-stakes examination of Arabic as a second language (ASL) or the high-stakes examination of English as a foreign language (EFL). The student questionnaires included both Likert-scale items and open-ended items.

In terms of the low-stakes ASL test, the sample comprised 62 student respondents from grades seven through nine. Eighty-six percent of the students said there were no special activities devoted to test preparation prior to the exam, and 72% said that no class time at all was devoted to the test. Fewer than two-thirds of the students were even aware of the test's existence (63% of the respondents), and 90% said they did not know what material is covered by the test. (These results concurred with data from the teachers, who said they do not inform their classes about the ASL test.) Nevertheless, 52% of the student respondents thought the test would affect their end-of-year course grades, and 62% said the test would influence "their knowledge of Arabic and future success in their studies" (ibid. p. 305). But 67% of the student respondents also said the test did not increase the prestige of Arabic as a school subject, and 65% said it was not important for them to succeed on this test (ibid.). Finally, 64% of the student respondents reported that the ASL test "does not reflect their true knowledge of Arabic" (ibid., p. 306).

In contrast, of the 50 student respondents to the questionnaire about the high-stakes EFL test, 54% of the higher level students reported "intense preparation for the exam" while only 13% of the lower level students (who would not take the test for some time) reported such preparation (Shohamy et al., 1996,

p. 308). When the new EFL test was put into place, 97% of the upper-level students reported being aware of the changes in the test (ibid., p. 309). In addition, 96% reported feeling "quite anxious about the test" (ibid., p. 310). Eighty-six percent of the students believed the EFL test could affect "their overall matriculation score to a large extent" (ibid.), and 70% of the student respondents believed that the results of the EFL exam could "affect their success in future studies" (ibid.). The student questionnaire results mirrored the teachers' views of the importance of this high-stakes exam: "82% of the students regard the exam as very important; and 84% of the students state that it is of considerable importance to them to succeed in the oral exam" (ibid., p. 311).

As Hughes (1993) has pointed out, the key question about the products of washback is whether or not it leads to learning (in our case, language learning). Shohamy et al. (1996) asked their student respondents whether and how the ASL test and the EFL test had promoted learning. With regard to the high-stakes EFL test, 68% of the students believed that the test promoted learning (from a large to a very large degree) and 92% said that the goal of the test was to promote learning. But in terms of the EFL test's impact on their *own* language learning, 46% believed that it had little or no impact while only 34% of the student respondents reported that "their command of English is affected to a large extent" (ibid., p. 312) by the test.

In terms of the low-stakes ASL test's impact on learning, Shohamy et al. (1996, p. 306) say only that "both teachers and students express negative feelings toward the test and complain that the test is of no importance and not essential in all course levels." An important issue that has not been investigated here is the extent to which the students' views are independent of or influenced by their teachers' views. Do teachers voice their opinions of tests to students, and if so, how are students influenced by their teachers' ideas?

Cheng (1997) reports on a study of language testing washback that she conducted at the time of a change in a major public examination: the Hong Kong Certificate of Education Examination (HKCEE). Her data included questionnaire responses from 42 students. The students' data revealed that they thought the HKCEE played "a 30% role in their learning," followed by the influence of future jobs, their parents' concerns, and competition with their classmates (ibid., p. 47). Thus the students' perception is that the exam is the single greatest factor influencing their English progress.

Alderson and Hamp-Lyons (1996), in a study of TOEFL preparation courses in the United States, interviewed students at three different institutions. Discussions with students in groups of 3 to 12 people were audio-recorded and transcribed. The language learners were asked for their ideas about how they would like TOEFL preparation classes to be conducted, compared to what they had already experienced. In the preliminary findings reported by Alderson and Hamp-Lyons, the students suggested

> having a placement test before a TOEFL preparation course . . . ; more opportunities for student participation and student questioning; diagnosis of individual student weaknesses; and the combination of self-study with revision in class . . . . The students also stressed the importance of practising English all the time. (1996, p. 285)

Alderson and Hamp-Lyons acknowledged, however, that their study would not be able to answer questions about the actual "effects of TOEFL on learners and learning" (ibid., p. 284).

In fact, only one of the language testing washback studies that I have found has documented any demonstrable gains in student learning that can be tied to the use of a test. Hughes (1988) was able to show that students' performance on the Michigan Test (a different, widely recognized measure of English proficiency) increased following the introduction of a new exam and subsequent changes in the English program (including the addition of summer courses in English) at a Turkish university:

> In previous years the number of students reaching the minimum requirement for their subject area had always been less than 50%. At the end of the academic year in which the new test was introduced, 72% of all [the program's] students achieved the minimum for liberal arts and education, a standard higher than for the other disciplines. At the end of the summer school the figure rose to 83%. This would seem to indicate a very great improvement over other years in the standard of English reached. (1988, p. 144)

Hughes also notes that "It was generally agreed in the University that there had been a marked improvement in the standard of English by comparison with previous years" (ibid., p. 145). Hughes' evidence includes data on both students' performance and people's perceptions of the students' gains, but this two-pronged approach is rare in the research literature on language testing washback. For this reason, we must keep in mind Messick's (1996) note of caution: "It is problematic to claim evidence of test washback if a logical or evidential link cannot be forged between the teaching or learning outcomes and the test properties thought to influence them" (p. 247).

Language Teachers and Washback

Program washback is one result of test-derived information obtained by someone professionally connected with a language program. These participants, as suggested above, could include anyone or a combination of the people who take significant roles in language instruction programs: teachers, counselors, administrators, course designers, materials developers, teacher supervisors, etc.

Of course, the most visible participants in program washback are language teachers. It is they who are the "front-line" conduits for the washback processes related to instruction. The importance of teachers in washback processes is emphasized by Alderson and Wall (1993, pp. 120-121) in several of their restatements of the washback hypothesis:

1. A test will influence teaching.

3. A test will influence what teachers teach; and

4. A test will influence how teachers teach.

7. A test will influence the rate and sequence of teaching; and

9. A test will influence the degree and depth of teaching; and

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

They also pose two contradictory hypotheses which demand investigation: (14) "Tests will have washback on all learners and teachers" and (15) "Tests will have washback effects for some learners and some teachers, but not for others" (ibid.).

The vast majority of the available empirical research on washback includes at least some focus on teachers. In fact, it is safe to say that teachers are the most frequently studied of all the participants in the washback process.

Shohamy notes the central role of teachers when she identifies some of the conditions that can lead to negative washback to programs (1992, p. 514; emphasis added):

> After all, when reliance is on tests to create change; when emphasis is mostly on proficiency and less on the means that lead to it (i.e., what takes place in the classroom as part of the learning process); when tests are introduced as authoritative tools, are judgmental, prescriptive, and dictated from above; *when the writing of tests does not involve those who are expected to carry out the change – the teachers;* and when the information tests provide is not detailed and specific and does not contain meaningful feedback and diagnosis that can be used for repair, it is difficult to expect that tests will lead to meaningful improvement in learning.

To counteract these potential negative washback effects, Shohamy argues for a continuous and cooperative feedback loop between external test developers and people working in the schools. (See also Shohamy, 1993b, p. 187)

In contrast to the perspective quoted above, Clark (1983) has commented on the potential beneficial washback of such proficiency tests. Again, he acknowledges the teacher's role:

> Equipped with an appropriate external-to-program indicator of acquired proficiency in the target language, referenced against the performance requirements inherent in real life language use, it becomes possible for a variety of individuals and groups, ranging from classroom teachers and their students, through local schools and school systems, to planners and implementers of broad-scale studies of the "national yield" of current language training – to determine the functional outcomes of instruction and to suggest possible further improvements in the instructional process on the basis of the information obtained. (p. 435)

Here Clark notes both the micro level and the macro level of washback (Bachman & Palmer, 1996, p. 30).

Hughes (1988) describes the reactions of Turkish university English teachers to the planned implementation of a new English test that had been based on a needs analysis of the learners' intended uses of English at the university:

> The first result of even threatening to introduce a test of this kind was to cause consternation amongst the [program's] teachers. They argued that their students could not possibly cope with such a test. Pointing out that the test would actually require the students to perform just the kind of tasks that they would meet in their first year as undergraduates (and thus the kind of task for which they, the teachers, had always been preparing them) was not very much appreciated. Many teachers were convinced that they were quite unable to provide the necessary training. (1988, p. 143)

This situation necessitated a number of changes in the program that included the introduction of a new syllabus and new textbooks, in addition to teacher training efforts. Hughes summarizes the effect on the teachers: "For the first time, at least in some years, the [program] teachers were compelled, by the test, to consider seriously just how to provide their students with training appropriate for the tasks which would face them at the end of the course" (ibid., p. 144).

Cheng's (1997) report on the revised Hong Kong Certificate of Education Examination (HKCEE) includes the analysis of questionnaires given to teachers, as well as observations of classroom lessons and of seminars conducted to advise teachers about the new exam. The seminars were offered by (1) the Hong Kong Examination Authority, (2) tertiary institutions, and (3) textbook publishers. In the results of the teachers' survey, "84% of the teachers commented that they would change their teaching methodology as a result of the introduction of the 1996 HKCEE" (ibid., p. 45).

However, Cheng notes that "what teachers stated they would like to change is not necessarily the same as what they actually would do in classrooms" (ibid., p. 49). For this reason, classroom observations should play an important role in any serious study of washback. Cheng was in an ideal situation to study washback via classroom observation, as indicated in the following details about her research design:

> Two cohorts of students [were] available for study: one group [which took] the old examination in 1995; the other [which took] the new exam in 1996. There [were] some teachers who [were] currently teaching both groups at the same time . . . (ibid., p. 44)

This context permitted Cheng to observe the same teachers teaching both groups of students (i.e., those studying on the old exam syllabus and those preparing for the new test).

One major change in the new version of the exam was that the previous sections on reading aloud in English were replaced by role-play tasks and group discussions. Cheng found in her classroom observations that under the old exam syllabus, teachers often had students practice reading aloud in choral groups. But under the new exam syllabus, "teachers no longer taught reading aloud. More and more time was spent on group discussions and oral presentations" (ibid., p. 48).

Cheng has also written a follow-up study which focuses specifically on how the revised exam influenced secondary school teachers in Hong Kong. She collected baseline data (before the exam revisions went into effect) by observing 15 lessons in six schools, and interviewing both teachers and students. Over a period of two years she then observed classes taught under the old and the new exam syllabuses. Three teachers were chosen as participants in a detailed case study. The observational data comprised videotapes and the researcher's fieldnotes. The videotapes were later analyzed using an instrument that targeted the categories of time, the focus of the lesson, activity types, interaction patterns, material used, and atmosphere (see Cheng, forthcoming, p. 13). Cheng discusses each teacher's behavior in detail but concludes that the interaction patterns had not changed much, that the lessons were conducted similarly before and after the introduction of the new exam syllabus, but that there were more varied teaching techniques observed after the implementation of the syllabus for the revised exam. The teachers reported that they did not understand what the new exam would be like or how they should teach under the new syllabus. Cheng notes that although the teachers relied heavily on practice examination papers, they also tried to use different teaching methods (e.g., one teacher began to use English-language newspapers in her classes).

Cheng comments on noticeable "differences between teachers, not within the teachers themselves" (ibid., p. 21) over time. Like other researchers in different contexts (see, e.g., Wall & Alderson, 1993; Watanabe, 1996), Cheng concludes that revised public exams "can to a large extent change the contents of teaching" (forthcoming, p. 22) but that in the case of the revised HKCEE very little changed with regard to interaction patterns, or the roles and functions of teachers and students.

Lam (1994) also investigated teachers' perceptions of changes brought about by the revisions in a national exam in Hong Kong (the Revised Use of English test). He focused specifically on what he called *methodology washback* – that is, trying to investigate *how* teachers teach English. Building on the multiple washback hypotheses posed by Alderson and Wall (1993), Lam wanted to investigate the possibility that the revised exam would "influence how teachers teach, i.e., the methodology and methods they use to prepare students for the public examination" (1994, p. 88). He surveyed 33 teachers who had taught under the syllabuses for both the old exam and the new exam, and 28 younger teachers who had taught only under the syllabus for the new exam. Among other things, the teachers who had worked under both systems were found to be "much more examination-oriented than their younger counterparts" (ibid., p. 91). Lam concludes that it is not sufficient to change exams: "The challenge is to change the teaching culture, to open teachers' eyes to the possibilities of exploiting the exam to achieve positive and worthwhile educational goals" (ibid., p. 96).

Andrews (1994b) also conducted questionnaire research involving teachers in the Hong Kong context. His approach was to survey the members of the working party that revised the exam, as well as secondary school teachers affected by the change. Thirty of these teachers had taught before the introduction of the oral component in the revised exam, and had thus had experience with both versions of the test, while 62 had not taught prior to the introduction of the oral component in the revised exam. Andrews found that both the teachers and the test developers emphasized teachers' willingness to devote time to improving students' speaking skills. However, the teachers felt that the impact of the new syllabus on the students' confidence and proficiency was not as strong as the test designers had hoped.

In general, the test designers placed greater emphasis than the teachers on the revised test's impact on students' motivation, confidence, and oral proficiency.

The Sri Lankan impact study (Wall & Alderson, 1993) is often cited as a landmark study in the investigation of washback. One of its key characteristics is the careful observation of teacher behavior. The situation allowed the researchers to compare data collected on several occasions, including baseline data collected before the new test was put in place and data collected after its implementation. Over a period of three years the authors and a team of local observers visited classrooms in five different areas of the country for six different rounds of observation, prior to and after the implementation of the new exam. Wall and Alderson found that before the exam was released

> those teachers who claimed to have knowledge of the exam could only give vague or confused explanations of what they expected, which showed that in reality they had no such knowledge; most teachers readily declared that they knew nothing. (ibid., p. 49)

In other words, prior to the implementation of the test, the teachers observed and interviewed were not sure what the new O-level exam would entail. For this reason, Wall and Alderson had confidence that the teaching they observed prior to the use of the new test was representative of instruction that had not been widely influenced by the new exam. Therefore, they had confidence later in attributing observed changes in teaching to the effect of the new exam.

Among many important results of the Sri Lankan impact study, Wall and Alderson make the following summary statements about the impact of the new Sri Lankan texts and tests on the teachers (ibid., p. 67):

1. A considerable number of teachers do not understand the philosophy/approach of the textbook. Many have not received adequate training and do not find that the *Teacher's Guides* on their own give enough guidance.

2. Many teachers are unable, or feel unable, to implement the recommended methodology. They either lack the skills or feel factors in their teaching situation prevent them from teaching the way they understood they should.

3. Many teachers are not aware of the nature of the exam – what is really being tested. They may never have received the official exam support documents or attended training sessions that would explain the skills students need to succeed at various exam tasks.

4. All teachers seem willing to go along with the demands of the exam (if only they knew what they were).

5. Many teachers are unable, or feel unable, to prepare their students for everything that might appear on the exam.

Wall and Alderson make an extremely important observation about the limitations of updated exams to bring about systemic change: "We now believe an exam on its own cannot reinforce an approach to teaching the educational system has not adequately prepared its teachers for" (ibid., p. 67). This last comment echoes Lam's point that having changed the exam, it is still a challenge to "change the teaching culture" (1994, p. 96).

In 1996, Wall revisited the Sri Lankan impact study. She succinctly states one of the key findings about teachers (1996, p. 348):

> The examination had had considerable impact on the content of English lessons and on the way teachers designed their classroom tests (some of this was positive and some negative), but it had had little to no impact on the methodology they used in the classroom or on the way they marked their pupils' test performance.

Wall goes on to outline possible reasons for the limited washback effect in some areas, drawing on innovation theory to provide a framework for analysis. (See, for example, Huberman & Miles, 1984; Kennedy, 1988; Markee, 1993; and Stoller, 1994.)

Andrews (1994a) also used innovation theory as a guiding framework when he reviewed the literature on the relationship between examinations and teachers' curricular innovations. He notes that there have been instances where efforts to make language teaching more communicative have been negatively influenced by "the perceived incompatibility of such an approach with prevailing examination practices" (p. 52). One important point he makes about the potential influence of exams to bring about (or prevent) curricular and methodological change is that "examination reform may indeed be a necessary condition for educational change; it is not, however, a sufficient condition" (ibid., p. 54). This point underscores Wall's and Alderson's idea that an exam on its own cannot bring about change if the educational system has not adequately prepared the teachers, and Lam's (1994, p. 96) point that it is a challenge to change the teaching culture, in spite of having changed the exam.

In Israel, Shohamy et al. (1996) observed the following results when the new test of Arabic (ASL) was originally instituted: teachers stopped teaching new material and began to review; textbooks were replaced with worksheets identical to the previous year's test; class activities became "test-like;" the classroom atmosphere became tense; and students and teachers were observed to be "highly motivated to master the materials" (ibid., p. 301). But, the authors note, "once the test had been administered, such teaching and learning activities ceased" (ibid.). In fact, their interview results showed that "once teachers learnt that the results had no personal immediate affect on them, they became relaxed and fearless, and thus the effect of the test decreased" (ibid., p. 314).

When the new EFL test was introduced in Israel, the teachers were observed to spend more class time teaching oral language, and the tasks and activities "were identical to those included in the test" (ibid., p. 301). Shohamy et al. (1996) noted significant differences between the experienced teachers and novice teachers. The former "turned to the test as their main source of guidance for teaching oral language and used only material to be included in the test" (ibid.), while the latter used "a variety of additional activities in the teaching of oral language" (ibid.).

In a completely different context (Japan) and using a different research design, Watanabe (1996) obtained results similar to those found by Wall and Alderson (1993) in Sri Lanka. He examined the classroom practice of two experienced male teachers, Teacher A and Teacher B, who were both observed teaching in two different English exam-preparation classes. One of each teacher's exam preparation classes was grammar-translation oriented and one was not. This design permitted Watanabe to compare not only two different types of courses, but also two different teachers working in the two contexts. Watanabe notes that "Teacher A appeared to be more [grammar translation] oriented than Teacher B, regardless of the type of course he was teaching" (1996, p. 327) and also that "there were differences between the two types of courses for Teacher B . . . but not for Teacher A" (ibid., p. 328).

Watanabe found that the presence of grammar translation questions on a particular university entrance exam did not influence these two teachers in the same way. He identified three possible factors that might promote or inhibit washback to the teachers: (1) the teachers' educational background and/or experiences; (2) differences in teachers' beliefs about effective teaching methods; and (3) the timing of the researcher's observations. (Teacher A was observed when the exams for which the students were preparing were six months away, while Teacher B was teaching exam-preparation classes just a month or so before the entrance examinations would occur.) Thus Watanabe concludes that "teacher factors may outweigh the influence of an examination" (ibid., p. 331) in terms of how exam preparation courses are actually taught.

A similar research design was used by Alderson and Hamp-Lyons (1996), who investigated TOEFL preparation courses at a language institute in the United States. They utilized three different types of data in this study: interviews with students in groups, interviews with teachers (both individuals and groups), and fieldnotes and audio-recordings made during classroom observations. Like Watanabe (1996), Alderson and Hamp-Lyons observed two different teachers while they taught both TOEFL preparation classes (a total of eight lessons) and other courses (including Structure, Listening, and Speaking, also for a total of eight lessons). This design permitted Alderson and Hamp-Lyons to compare the TOEFL preparation and non-TOEFL preparation classes, as well as the two teachers' behaviors in both types of classes.

Among their findings about the TOEFL and non-TOEFL oriented classes, Alderson and Hamp-Lyons list the following (1996, pp. 288-289):

1. Test-taking is much more common in TOEFL classes;

2. teachers talk more and students have less time available to talk in TOEFL classes;

3. there is less turn-taking and turns are somewhat longer in TOEFL classes;

4. much less time is spent on pairwork [in TOEFL classes];

5. the TOEFL is referred to much more in TOEFL classes;

6. metalanguage is used much more in TOEFL classes;

7. TOEFL classes are somewhat more routinized; and

8. there is much more laughter in non-TOEFL classes.

Many of these findings are not surprising. However, after subsequently analysing the teachers' behavior, Alderson and Hamp-Lyons note that "the differences between the two teachers are at least as great as the differences between TOEFL and non-TOEFL classes" (ibid., p. 290).

In considering the varied research about washback and language teachers, we can see that teachers' classroom behavior can either support or override the intended positive washback effect of new or revised tests. There have also been differences observed between novice and experienced teachers with respect to washback. We have seen that in many contexts teachers change the content of their teaching but not their methods as a result of examination changes. We now turn to the available empirical evidence on people other than teachers and learners who are also cited in the literature on washback.

Other Participants in Washback Processes

The research on other parties who try to create, or are influenced by, program washback is less widely developed than the research on language learners and teachers. The other participants can include test developers (Andrews, 1994b; Andrews & Fullilove, 1994), teacher educators and curriculum planners (Andrews & Fullilove, 1994), teacher advisors (Wall & Alderson, 1993), principals and other administrators (Fullilove, 1992; Hughes, 1993; Shohamy, 1993b; Shohamy, Donitsa-Schmidt, & Ferman, 1996), language inspectors (Shohamy, Donitsa-Schmidt, & Ferman, 1996), end-users (Andrews & Fullilove, 1994), materials developers and publishers (Cheng, 1997; Hughes, 1993), and even parents (Andrews, 1994a; Cheng, 1997; Fullilove, 1992; Ingulsrud, 1994; Shohamy, Donitsa-Schmidt, & Ferman, 1996).

A repeated theme found in the literature on these other participants, particularly test designers and policy makers, is the dynamic tension between (1) the intended positive washback in implementing new or revised exams and (2) how that impact is realized in classroom practices. Andrews and Fullilove (1994, pp. 57-58) assert that in cases where new or revised tests have a negative washback effect,

> the reforms in language teaching proposed by teacher educators and curriculum planners have been undermined by the conflicting message implicit in the tests, especially in those countries where examinations are highly important and yet where the examination format has been particularly resistant to change.

Likewise Cheng (forthcoming, p. 9) summarizes her review of the literature on first language testing washback by noting that while policy makers and test developers may hold a particular view of the desired changes a new or revised test will promote, these images "are translated imperfectly by practitioners."

There are often interesting discrepancies in the perspectives of various participants in the washback process. For example, according to Shohamy et al. (1996), the purposes of the Arabic as a second language (ASL) test introduced in Israel in 1988 were to

> raise the prestige of the Arabic language, to equalize levels of teaching in schools, to force teachers to increase the rate of teaching (especially the alphabet), and to increase the motivation of both teachers and students to teach and learn Arabic. (ibid., p. 301)

The data in their study included structured interviews with teachers and with inspectors of ASL (ibid., p. 302). The authors found in their interviews that the inspectors were aware of high test anxiety (among both teachers and students) in previous years' tests, but that the test anxiety had decreased and that some teachers did not even administer the test. Others treated it as a quiz that required no preparation. However, Shohamy et al. stated the inspectors felt that

> it is essential that the test continue to be administered as they believe that there would be a major and significant drop in the level of Arabic proficiency in the country were the test to be cancelled. Moreover, the Inspectorate claims that there would be a decrease in the number of students studying Arabic since the test promotes the status of Arabic as perceived by teachers, students and parents. (ibid.)

This finding illustrates the disparate views held by the inspectors, on the one hand, and the students and teachers of Arabic on the other.

When Shohamy et al. (1996) interviewed the inspectors associated with the high-stakes EFL exam, they found that "the Inspectors claim that the introduction of the oral test has had a very positive educational impact and the washback on teaching has been tremendous" (ibid., p. 312). The inspectors also feel that the test has successfully promoted learning, particularly of oral skills. They believe that "were the oral exam to be cancelled, teachers would cease teaching oral proficiency" (ibid.). In other words, in both cases, the inspectors of the Arabic and English exams see their respective tests as "necessary, important and effective" (ibid., p. 313). However, Shohamy et al. point out that this position "is in contrast to how teachers and students perceive the test" (ibid.) and that in general "unlike teachers and students, the bureaucrats portray a much more positive picture" (ibid.).

Another set of participants who may be influenced by or try to utilize washback is the "end-users" – that is, people who, in the future of the language learners, will in some way benefit from their target language proficiency. (In the English for Specific Purposes [ESP] literature, the students' future employers are often the end-users.) In the case of the Hong Kong Use of English test, Andrews and Fullilove (1994) note that the addition of an oral component was promoted by colleges and universities, which generated program washback to the secondary schools:

> Tertiary institutions urged the [Hong Kong Examination Authority] to add an oral component to the examination for two reasons: (1) to provide the universities with more information about their potential students' ability to communicate orally; and (2) to try

> and improve the oral proficiency of tertiary students generally by encouraging the teaching and practice of oral skills. (ibid., p. 62)

Point (2) directly hinges on the assumption of program washback to the secondary schools. In this case the tertiary institutions may be seen as the "end-users" who have a stake in the product of secondary school English teaching – that is, the future university students' ability to use oral English.

Finally, parents are occasionally included in research on washback phenomena. Andrews (1994a) notes that there is "widespread acceptance of the assertion that tests, especially public examinations, exert an influence on teachers, learners *and parents*" (p. 45; emphasis added). For instance, after asserting (and providing convincing evidence) that "Hong Kong is an examination-mad town" (1992, p. 131), Fullilove explains,

> Given this orientation of Hong Kong society, it is little wonder that the local public examinations create extreme pressures on Hong Kong students from the day they first enter the educational system – or more accurately, before that day, when anxious parents take their tiny 'scholars' to pre-kindergarten interviews to gain admission to choice places even on this lowest rung of the educational ladder. (ibid.)

However, there is relatively little research that documents the parents' own perceptions of language testing washback. The studies that document parents' ideas typically do so through the students' perspective.

For instance, Shohamy et al. (1996, p. 304) found in their Israeli student questionnaire data that 77% of the student respondents said that their parents do not know about the low-stakes ASL test. In contrast, the EFL student respondents' data indicated that "60% of their parents [were] aware of the change in the new EFL test" (ibid., p. 309). Cheng reports that secondary school students in Hong Kong said the exam results influenced their parents in the following ways: "(1) the advice their parents gave them, (2) parents became tense and anxious, and (3) parents put more pressure on them" (1997, p. 47). In discussing university entrance examinations in Japan, Ingulsrud states that "If parents are serious about their children entering a prestigious university, they will send them to after-school and holiday coaching schools" (1994, p. 71).

We can see from this review of the available empirical research that teachers are the most frequently studied participants in washback processes. However, many other people are also involved in language testing washback. The comparative dearth of empirical findings on students suggests that more research is needed about how tests actually influence second language learners' behavior and attitudes.

Part IV: Research on Processes and Products of Washback

The processes posited by Hughes (1993) and represented only by arrows in Figure 1 are the least tangible aspects of the washback phenomenon. Hughes defined *processes* as "any actions taken by the participants which may contribute to the process of learning" (1993, p. 2). He included materials development, changes in teaching methods or content, syllabus design, the learners' use of learning strategies or test-taking strategies, etc. as examples of the processes by which washback occurs.

Hughes originally defined the *products* associated with washback as "what is learned (facts, skills, etc.) and the quality of learning (fluency, etc.)" (1993, p. 2). We encounter problems, however, in trying to untangle participants and processes from products in the available research literature. Much of the research cited above about teachers and washback describes the various processes teachers use to try to increase students' mastery of skills and/or their test scores. Such processes include reviewing previous years' test papers, or "seen passages" as they are called (Wall & Alderson, 1993, p. 63), and teaching additional classes (ibid., p. 64). Shohamy (1993b, p. 186) highlights processes as well when she claims that negative washback often brings about "underemphasis on the means by which the learner arrives at proficiency." These means include both processes and products: "instructional activities, teaching methods, classroom learning, curricula, and textbooks" (ibid.).

Score Reporting

One washback process that has received some speculative attention, but has not yet prompted empirical research, is the effect of various approaches to score reporting (Bailey, 1996a, pp. 271-272). Direct feedback (albeit sometimes limited) about a learner's actual performance on a test is provided in the score report. Such reports are usually sent to the learners themselves, perhaps to their sponsors, and to those they designate (e.g., admissions officers of the programs that they wish to enter). In the case of the Institutional TOEFL or other institutionally administered exams, such as the Speaking Proficiency English Assessment Kit (SPEAK®), the results will also be known by the program staff who administer and score the test. Typically, such score reports are brief and to the point. For many years, for example, TOEFL score reports have given a total score and three subtest scores: Listening, Structure and Written Expression, and Reading and Vocabulary.

The General Test of English Language Proficiency, or G-TELP®, a commercially developed criterion-referenced test of English proficiency, provides somewhat more detailed information to the test-takers in its score report. The *G-TELP Information Bulletin* (1990, p. 19) explains the intended use of the score report:

> The G-TELP Test Score Report provides a level of mastery score and two profiles of the test-taker's performance. These indicate the test-taker's degree of mastery of the language tasks for the proficiency level at which he/she was tested. The Test Score Report, used in conjunction with the G-TELP Test Descriptors, . . . [is] intended to aid the test-taker and score user in interpreting the test-taker's performance.

The G-TELP Score Report provides the test-takers not only with information about their performance on the subtests (Listening, Reading and Vocabulary, and Grammar), but also with specific feedback in percentage terms about their performance on the tasks and structures assessed in these subtests. For

example, the advertising bulletin put out by the G-TELP Committee of Korea (undated but obtained in October, 1996) shows and explains a sample score report for overall proficiency and another for the speaking test. The latter provides individual students with percentage level scores for specific functional tasks they complete during the test (e.g., giving personal information, narrating a story from pictures, expressing and supporting an opinion, giving directions from a map, etc.).

Alderson, Clapham, and Wall devote an entire chapter to "Reporting Scores and Setting Pass Marks" and another to "Post-test Reports" (1995, pp. 148-169 and pp. 197-217, respectively). They note that "profiling of test results is considered by many to be superior to reporting one overall result" (ibid., p. 155) because different individuals may achieve the same overall score through many different possible combinations of subtest scores. A profile provides more information to both the test candidates and the end users. In discussing the value of post-test reports, Alderson et al. state that:

> Tests can have important consequences for those who take them, and for those who use the results. It is therefore incumbent upon those who produce the tests to provide all the evidence they can muster for the validity, reliability, and meaningfulness of their tests and the results. (ibid., p. 197)

They point out that one obvious audience for post-test reports would be the teachers who "will be preparing other students to take the test in the future" (ibid.). Specifically, Alderson et al. suggests that teachers need "summaries of the types of problems candidates experienced on different parts of the test and advice about how to prepare candidates more effectively in the future" (ibid., p. 203). They continue, "Since one common way of preparing students for tests is by using past papers, it is important for teachers to know whether answers that their students propose would have been considered 'acceptable' by the testing body" (ibid., p. 204). Thus post-test reports can potentially serve an important function in the washback process.

Wesche (1983) connects the type of score report to the nature of the decision to be made with the test results: "If the purpose of testing is diagnostic or to evaluate progress in a language training program, detailed scoring grids might be in order" (p. 47). But she also notes that "global native-speaker judgments of whether or not the learner has the requisite second language communication skills might be more appropriate for placement or entrance purposes" (ibid.).

Spolsky (1990) has tied the use of detailed score reports directly to washback:

> Because there is a natural tendency on the part of those who use test results to take shortcuts, there is a moral responsibility on testers to see that results are not just accurate but do not lend themselves to too-quick interpretation. For this purpose, profiles rather than single scores seem of special value: score reports that include several skills, tested in different ways, and adding if at all possible some time dimension. (p. 12)

Spolsky goes on to explain the significance of this time dimension:

> This last factor is most seldom included, but can be most revealing. Proficiency tests usually ignore the dynamic dimension; they give no way of deciding whether the subject is in the process of rapid language learning or has long since reached a plateau. (ibid.)

This same issue has been discussed by Haas (1990), a university admissions officer. He notes that "language skills are not static: they can change considerably" in just one year (ibid., p. 11). Therefore, he suggests admissions officers, administrators, and teachers must keep in mind that score reports are often several months old when they are used for decision-making purposes.

Chapelle and Douglas (1993) discuss score reporting in computer-adaptive testing of communicative competence. The information from such a test would not be "a single score on a unidimensional scale; instead, information on multiple competencies would be reported and used diagnostically as needed" (p. 9).

Writing in a similar vein, Shohamy (1993b) emphasizes the importance of detailed score reporting. She notes that

> the only test scores that will be used are those that provide new and meaningful information. It needs to be detailed, descriptive, and diagnostic, able to address a variety of dimensions, and not collapsible into one general score. (p. 189)

Shohamy adds that if exam results are expected to bring about change, such results must be "translated into instructional activities and actual strategies for teaching and learning" (ibid.). She describes a diagnostic feedback model of teaching and testing (using Hebrew as the target language), in which test developers and school personnel collaborate to help students improve their learning, and subsequently their test performance. Five different kinds of analyses are conducted on the test data: general, diagnostic, comparative, qualitative, and itemized. Shohamy states that "the five types of results are presented along with visuals and graphics so that school personnel with no background in statistics can understand them" (ibid., pp. 198-199).

Although the authors cited above have discussed score reporting, I have located no empirical language testing research connecting the score reporting process to demonstrable learner washback leading to improved performance. Presumably official score reports provide some feedback to the test-takers, but their utility and their impact depend on several factors. For example, scores should be accurately reported in a timely manner. The information should be clearly presented in non-technical terms that are interpretable by laypersons. The information should be detailed enough to allow test-takers and their teachers to plan a course of action for improving the learners' target language skills. Clearly, further investigation into the effects of score reporting is needed.

Washback on Materials for Language Teaching and Learning

In October 1996, I visited Korea, a country that in recent years, has become a major consumer of the Test of English for International Communication (TOEIC®), a product of Educational Testing Service (ETS). In a subway station in Seoul, I took a photograph of a large billboard advertising TOEIC test preparation materials. The sign pictures a stack of 22 different TOEIC-oriented books and another group of 10 other products, including what look like boxes of computer diskettes. Although I could not read the Korean text, the English parts of the sign listed "Mini TOEIC, Number One TOEIC, Super Elite TOEIC, Elite TOEIC, Speedup TOEIC, TOEIC Program, TOEIC Academy, TOEIC Academy Plus, and TOEIC Bible" – the names of the products. Someone must be buying these books!

TOEFL test preparation materials are also numerous and widely available. Pierce points out that such materials are indirect evidence of washback. She states that reports of TOEFL test washback "remain anecdotal" and that its existence "can only be extrapolated from the vibrant industry in TOEFL preparation books" (1992, p. 687).

Publishers are participants in language testing washback through three related but distinct processes. First, there are several major publishers who actually produce tests or publish tests written by people in the language teaching field. Second, publishers produce subject matter coursebooks that may influence or be influenced by exams. Third, they publish textbooks that are designed explicitly as test preparation materials. Hilke and Wadden refer to this side of the publishing industry as "a quasi-educational business . . . which churns out scores of preparatory guides, seminars, and sample tests in countries around the globe" (1996, p. 53).

This section of the monograph reviews the findings of the available research literature on washback and materials development. As noted above, *textbook washback* has specifically been identified as one possible result of test use (Lam, 1994, p. 85). This effect may be particularly important, or at least most noticeable, in the context of a change in an exam or the implementation of a new exam.

In the Sri Lankan situation, new textbooks had been introduced and "the exam was intended to reinforce the textbook" (Wall & Alderson, 1993, p. 43). Wall and Alderson posed the following important questions about the washback effect, with reference to textbooks:

> Do the teachers understand the approach of the textbook? Are they prepared to accept this? Are they able to implement the new ideas? Are they aware of the nature of the exam? Are they willing to go along with its demands? Are they able to prepare their students for what is to come? (ibid., p. 48)

Wall and Alderson were careful, in the Sri Lankan impact study observations, to tease out the effect of the textbook series from the effects of the examination based on the texts. They concluded, following observations and interviews with teachers, that a large number of teachers did not understand the approach or philosophy embodied in the texts (1993, p. 67) and were unaware of the nature of the new exam (ibid.). They also found that "many teachers believe they have to follow the textbook faithfully because the exam may test any of the content therein" (ibid., p. 63).

Hong Kong is another situation where textbook washback is particularly evident. For instance, in 1981 Johnson and Wong said of the public examinations faced by secondary school students, "We have a stable but stultifying interdependence of language teaching, language testing, and instructional materials within which we are presently trapped" (1981, p. 285). Fullilove (1992), writing about the oral component of the Revised Use of English (RUE) examination in Hong Kong, made the following prediction:

> Presumably this new exam will lead to the publication of new textbooks with accompanying audiotapes and videotapes in Hong Kong directed towards the improvement of speaking in general and the practicing of the format of the new exam in particular. (p. 144)

Lam explains, however, that the RUE test in Hong Kong was meant to be neither textbook-based nor content-based (1994, p. 88). Instead, it was intended to be a proficiency test. Proficiency tests are not tied to particular materials or programs of instruction. They are meant to assess students' general levels of language ability. (See Bachman, 1990; Brown, 1996; Finnochiaro & Sako, 1983; Henning, 1987; Hughes, 1989; Lowe, 1988; Oller, 1979; and Shohamy, 1992.) Lam (1994) notes that the aim of the RUE was "to foster the development of students' English language skills in order to equip them for tertiary education and/or employment" (p. 88). Nevertheless, based on self-report data from teachers, Lam acknowledges (ibid., p. 90) that a textbook-reliant methodology in Hong Kong is still very significant, i.e. about 50% of the teachers appear to be "textbook slaves" in teaching the sections of the test related to listening, reading, and language systems, and practical skills for work and study. (The teachers reported themselves as being less reliant on textbooks in preparing students for the writing and oral sections of the test.) Lam feels that this reliance on textbooks in this context is evidence of

> negative washback because instead of introducing more authentic materials [the teachers] prefer to use commercial textbooks, most of which are basically modified copies of the exam paper. (ibid.)

In a context like Hong Kong, where the exam in question has very high stakes (at both the micro and the macro level) and where so many English teachers are not native speakers of the target language, it is not surprising to find this reliance on textbooks in preparing students for the exam.

Fullilove (1992) also discusses teaching materials designed to prepare students for the public examinations. He states that many such texts are "little more than clones of past exam papers" (p. 139). The educational result is that "some students in Hong Kong, particularly the weaker candidates, tend to spend long hours memorising those model answers, rather . . . than actually learning how to answer similar questions" (ibid.).

Cheng (1997) also comments on textbook washback in the Hong Kong context. She describes the commercial and pedagogic effects of the Hong Kong Certificate of Education Examination (HKCEE) revision in this way:

> By the time the examination syllabus affected teaching in Hong Kong secondary schools, nearly every school had changed . . . textbooks for the students. Almost all textbooks are labelled specifically "For the New Certificate Syllabus" . . . Publishers in Hong Kong worked really hard and quickly to get textbooks ready for the schools. The main reason for this might be the way in which Hong Kong society develops quickly, especially in commercial matters. (ibid., p. 50)

Cheng goes on to explain that textbooks are the most direct form of teaching support available in Hong Kong. Publishers provide teaching materials, detailed methods for conducting suggested activities, advice about time distribution, and seminars for using their materials. In an interview a textbook publisher told Cheng,

> Anyone who speaks some English would be able to teach English in Hong Kong as we have provided everything for them . . . . Sometimes teachers phone us when they come across difficulties in teaching a particular unit or task. And we would write a detailed plan for them . . . (ibid.)

This reliance on textbooks is an important contextual variable to be understood in any investigation of washback in Hong Kong. Cheng explains that teaching and learning in this region are usually based on a major textbook and an accompanying set of workbooks:

> These workbooks are specifically designed to prepare students for specific examination papers in the HKCEE. Therefore, it could be assumed if teachers rely on textbooks a lot (which was evident through school visits), and if textbooks catering for the 1996 HKCEE have really integrated the underlying theory behind the change of this public examination and realized this through the language activities in their textbooks, it would be likely that the 1996 HKCEE would have certain washback effects on the teaching and learning of English in Hong Kong schools, given the importance of this public examination. (ibid., p. 50)

Thus Cheng has identified a key issue in the interface of processes and products related to language testing washback. A first-phase process is that new texts are generated by materials writers and publishers, and a second-phase process is that teachers use the texts to teach students. Two questions arise, however: first, do the texts correctly embody the constructs underlying the exam, and second, do teachers understand and convey that information to students?

In considering the important role textbooks play in the processes of washback, we need to keep in mind a caveat raised by Andrews. That is, while exam-related textbooks may be designed based on information supplied by the examining body responsible for an innovation in assessment (whether it be the Ministry of Education, a university-based test development team, or a pair of teachers preparing TOEFL test preparation materials to submit for publication), ". . . the final product will not be moulded according to the innovators' view of what is desirable in terms of teaching, but rather according to the publishers' view of what will sell" (1994b, pp. 79-80). Andrews notes that in the Hong Kong context this tendency leads to examination-specific materials that limit the focus of teachers and learners – a

problem that has been referred to as "narrowing of the curriculum" (see, e.g., Raimes, 1990; Shohamy, 1992, 1993a; and Wall & Alderson, 1993). Shohamy (1992) states that negative washback to programs can result in "the narrowing of the curriculum in ways inconsistent with real learning and the real needs of . . . students" (p. 514).

Cheng would agree with Andrews' concerns about the narrowing of the curriculum through textbooks. Based on her classroom observations she notes that teachers did adopt different classroom behaviors after the implementation of a major revision of HKCEE:

> However, these obvious changes made in teaching lay in the different activities designed in the textbooks the teachers employed. This means the textbooks changed [the teachers'] ways of organizing classroom activities according to the textbook publishers' understanding of the 1996 HKCEE. Those activities or tasks are designed only on the basis of the limited sample exam formats provided by the [Hong Kong Examination Authority] . . . (Cheng, 1997, p. 51)

Cheng concludes that, at least at this stage of the implementation of a new syllabus that was designed to prepare students for the revised HKCEE, "In effect, teachers follow the new syllabus simply by adherence to the new textbooks" (ibid.).

The potential for textbooks to lead to negative washback has also been documented elsewhere. For instance, Shohamy, Donitsa-Schmidt, and Ferman (1996) describe a reading comprehension test for middle school children that had been implemented in Israel but then discontinued for a variety of reasons. One reason was that "mass preparation for the test in the form of texts and tasks identical to those used on the test became the sole teaching material in classes" (p. 300). Thus although the Israeli context is quite different from that investigated in Hong Kong by Andrews (1994a; 1994b), Cheng (1997), and Lam (1994), the materials-oriented products and processes related to washback are quite similar.

Shohamy, Donitsa-Schmidt, and Ferman (1996) also report on the courseware available to support the recently introduced high-stakes EFL exam in Israel. The materials include a videocassette and book for teaching literature, a TV series about extended speech, an audioseries about the kinds of literary items which appear on the exam, newspapers to help students prepare for the extended interview component, and cue-cards to teach them the kinds of question-formation expected in the role-play tasks (ibid., p. 309). In contrast, for the less successful, low-stakes test of Arabic as a second language (ASL), Shohamy et al. report that "no special courseware for the ASL test has been generated since 1993" (ibid., p. 304).

Teachers' reliance on textbooks in preparing students for tests is not limited to Asia, Israel, or developing nations, however. In their study of TOEFL test preparation classes in the United States, Alderson and Hamp-Lyons (1996) observed classes and interviewed both teachers and students. They comment,

> In our discussions with teachers we found little serious consideration of what might be an appropriate or successful way to teach TOEFL; most teachers just seemed to do what the book says and what they claim the students want. (p. 286)

Alderson and Hamp-Lyons put the teachers' practices and attitudes about such materials into a broader context. They stress that the teachers "cannot be considered to be at fault in this: they are merely part of a huge test preparation industry fueled by students' anxiety to succeed on the high-stakes test" (ibid., p. 293).

As a follow-up to the Alderson and Hamp-Lyons study, Hamp-Lyons (1996) used a continuum of criteria (influenced by Mehrens and Kaminsky [1989] and Popham [1991]) to evaluate test preparation materials. She analyzed five widely used TOEFL test preparation texts (Broukal & Nolan-Woods, 1995; Lougheed, 1992; Rogers, 1993; Schutz, Derwing, Palmer, & Steed, 1991; and Sharpe, 1996). Hamp-Lyons (1996) notes that such textbooks

> play a major part in the TOEFL test preparation industry and may form the largest market for EFL, and to a lesser extent ESL, textbooks. They have, then, tremendous potential washback effect on teaching and learning. (p. 3)

Hamp-Lyons continues (ibid., p. 6), "These test preparation books consist, to a greater or lesser extent, of practice tests or 'exercises' which are themselves of exactly the same format as the subsection of the test they are preparing for." She complains that although such texts do prepare students to some extent before the test, the texts "typically contain *no* material to help teachers or students *after* they have taken the test or a practice test" (ibid.).

In a study similar to Hamp-Lyons' analysis of TOEFL test preparation materials, Hilke and Wadden (1996) analyzed the content and item focus of recent versions of the TOEFL test. They then analyzed the practice tests available in 10 different commercially produced TOEFL test preparation books that are widely used in Japan, including some of those that were reviewed by Hamp-Lyons (1996). Their study is based on the assumption that

> *it is* possible to effectively prepare for the TOEFL but that the effectiveness of that preparation will be relative, at least in part, to how accurately the preparation materials reflect the exam itself. (Hilke & Wadden, 1996, p. 53)

Even though these authors found wide variability in the extent to which the commercially published practice tests paralleled the actual TOEFL test forms they examined, they note that the test preparation textbooks "typically claimed – implicitly and explicitly – to reliably represent the exam" (ibid., p. 56).

One new TOEFL test preparation package recently released in the United States reflects the emphasis publishers put on preparing students for new or revised exams (American Language Academy, 1997). The advertisement states,

> TOEFL CBT (Computer-Based Testing) is coming. *TOEFL Mastery* is here *now*. Your students can be using *TOEFL Mastery* in minutes to improve and predict their TOEFL scores and to strengthen their English language skills. *TOEFL Mastery* is faster, easier, more flexible, and more fun to use than traditional paper and cassette courses. *TOEFL Mastery* mimics the actual test including sound, timed tests, and questions just like the ones in the TOEFL test. A review mode gives students a chance to look at their mistakes and receive valuable context-sensitive feedback for guidance and suggestions for making the right choice the next time.

This advertisement emphasizes the product's parallel focus on test preparation and language learning.

The advertisement suggests that the TOEFL 2000 project is viewed by the field as revolutionary – at least in its delivery mode. If the test delivery system has any washback whatsoever, then we would expect to see at least two widescale responses to the TOEFL 2000 project being a computer-based test. First, commercially developed TOEFL test preparation materials, like the package described above, should immediately incorporate the computer-based delivery system (and as a corollary, TOEFL test preparation materials based on the paper-and-pencil version of the test will either be revised or discontinued). Second, intensive English programs and other programs that offer TOEFL test preparation courses should quickly change their curricula to reflect the computer-based system. This response might include buying computer equipment if it had not already been a part of the program's assets.

In this section of the monograph we have considered the processes and products posited in Figure 1, and found them to be largely intertwined. Although relatively little has been written about score reporting, a great deal of information has emerged about washback and language teaching materials, especially test preparation materials. Based on this literature review, it is not an exaggeration to say that the role of textbooks, and of the authors and publishers who produce them, is crucial in the washback process. Much more research is needed in this area.

Part V: Investigating Washback from the TOEFL 2000

The introduction of the TOEFL 2000, the first internationally administered computer-based test of English proficiency, presents numerous opportunities for investing language testing washback. However, important issues of both research methodology and research foci should be considered before undertaking such investigations.

Research Methods Issues

In this section of the monograph I will discuss research methods issues which are particularly important in investigations of washback. These issues include observing classrooms and asking participants about washback, various kinds of triangulation, and the use of quantitative and qualitative data.

Watching and Asking. If the core of any definition of washback has to do with the effects of tests on learning and teaching, then it is necessary to document those effects – both by asking about and by watching teaching and learning.

Observation has long been accepted as an important feature in language teacher education and supervision, but in the past two decades it has become established as the key process in language classroom research as well. (See Allwright, 1988; Allwright and Bailey, 1991.) Whether it involves audiotaping, videotaping, taking fieldnotes, using a coding schedule, or any combination of these data collection procedures, *observation* is defined here as the systematic, purposeful recording of interactions and events in classrooms. Although some studies of washback have depended only on interviews or questionnaires, much of the recent research reviewed in Part III and Part IV of this monograph included a strong observational component in the data collection phase. I would argue that the observational component is necessary in order to understand washback.

In a study of first-language education that did not include an observational component, Herman and Golan (1993) conducted survey research among matched pairs of teachers from two different kinds of schools: those where test scores had increased, and those where test scores had decreased or remained the same. Among other things, the teachers' self-report responses to Herman and Golan's questionnaire yielded the following findings:

1. Teachers feel pressure to improve students' test scores.

2. Testing affects instructional planning and delivery.

3. Substantial time is spent preparing students for testing.

4. Non-tested subjects also get some attention.

Without observational data, however, we do not know how such pressure influences teaching, in what ways tests influence planning and delivery, how much time is spent preparing students for testing, and what kind of attention is given to those subject areas that are not covered in the tests. Thus, survey data alone are useful but insufficient for understanding language testing washback.

Wall and Alderson's (1993) commentary on the different types of data they collected are particularly illuminating in this regard. They note that the observational data were important, but did not tell them

> *why* the teachers do what they do, what they understand about the underlying principles of the textbook and the examination, and what they believe to be effective means of teaching and learning. Observations on their own cannot give a full account of what is happening in classrooms. (ibid., p. 62)

Indeed, their observational data raised many questions for Wall and Alderson which could only be answered through more direct access to the teachers' thought processes. They continue,

> It was important for us to complement the classroom observations with teacher interviews, questionnaires to teachers and teacher advisers, and analyses of materials (especially tests) teachers had prepared for classes. (ibid., p. 63)

But Wall and Alderson also argue strongly for the inclusion of observational data because, they say,

> without them we would not have known that certain questions needed to be asked, . . . we might not have been able to understand some of the answers teachers gave us in the questionnaires and interviews, . . . [and] we would have had no choice but to believe what the teachers told us. (ibid., pp. 64-65)

Allwright and Bailey (1991, pp. 3-4) note that collecting data in classroom research consists basically of watching or asking. Wall and Alderson (1993) explain why both approaches are crucial in investigations of washback:

> Above all, we would not have known that the exam had virtually no impact on methodology if we had not observed classes. However, we would not have been able to understand *why* the exam had no impact on how teachers taught without discussions with teachers *after* having observed their classes. (p. 65)

They conclude that "observations and interviews, questionnaires, [and] discussions necessarily complement each other in studies of this type" (ibid.). Wall's and Alderson's emphasis on *why* the teachers responded to the new test as they did suggests that future studies of washback could benefit from the findings and the research procedures, such as stimulated recall (see, e.g., Bailey, 1996b; Johnson, 1992; and Nunan, 1996b), of teacher cognition studies. (Freeman [1996] provides a helpful overview.)

In the washback literature reviewed in this manuscript, no single uniform questionnaire has emerged as being widely used to survey students or teachers about language testing washback. Perhaps this is because there are advantages to surveying students in their native language, or because such studies have tended to focus on the washback from a particular exam that is used locally, even if at the national level. This was certainly the case with the research on the needs-based exam at a Turkish university (Hughes, 1988), the new exams in Sri Lanka (Pearson, 1988; Wall, 1996; Wall and Alderson, 1993), the

Japanese university entrance exams (Buck, 1988; Ingulsrud, 1994; Watanabe, 1996), and the EFL and ASL exams in Israel (Shohamy, 1992, 1993a, 1993b; Shohamy et al., 1986; and Shohamy et al., 1996), as well as all the research on the various exam revisions in Hong Kong (Andrews, 1994a, 1994b; Andrews & Fullilove, 1994; Cheng, 1997, forthcoming; Fullilove, 1992; Lam, 1994; Smallwood, 1994). With the appearance of the TOEFL 2000 we will have the opportunity to investigate washback on an international basis, with a population that includes varied first-language backgrounds, as well as many different cultures and national origins. For this reason, the systematic development of a widely usable questionnaire for teachers and another for students would be a valuable contribution to the available methodological tools for investigating washback.

Triangulation. Anthropologists, and more recently applied linguists, have borrowed the concept of *triangulation* from navigation and land surveying. Hammersley and Atkinson (1983) explain the idea this way:

> The term *triangulation* derives from a loose analogy with navigation and surveying. For someone wanting to locate their position on a map, a single landmark can only provide the information that they are situated somewhere along a line in a particular direction from the landmark. With two landmarks, however, their exact position can be pinpointed by taking bearings on both landmarks; they are at the point where the two lines cross. (p. 198)

Hammersley and Atkinson make the analogy that in social research, the error inherent in one type or source of data may go undetected, but if "different types of data lead to the same conclusion, one can be a little more confident in that conclusion" (ibid.). Various approaches to triangulation may be employed to increase the quality control and representativeness of a study.

There are essentially four types of triangulation. The first is *data triangulation*, in which data from more than one source are brought to bear in answering a research question (e.g., the data from teachers, language learners, and inspectors in the study by Shohamy et al., 1996). Second, *investigator* (or *researcher*) *triangulation* refers to using more than one person to collect and/or analyze the data. In *theory triangulation* more than one theory is used to generate the research questions and/or interpret the findings. Finally, in *methodological* (or *technique*) *triangulation* more than one procedure is used for eliciting data – for instance, Wall and Alderson's (1993) use of interviews and classroom observations. (See Allwright & Bailey, 1991, p. 73; Denzin, 1970, p. 472; and van Lier, 1988, p. 13 for more information about triangulation.)

Wall and Alderson's (1993) study in Sri Lanka provides an excellent example of investigator triangulation and methodological triangulation. Investigator triangulation is illustrated by the fact that "seven Sri Lankan teachers based in five different parts of the country agreed to act as observers" (p. 49) and went through a three-month training program to prepare for this role. (Later one of the original observers was replaced by "several new team members" [ibid., p. 50].) The observers sent their data to England, "where a member of the Lancaster team would analyze their data . . . .The analyst would then send feedback to the observers and instructions for the next round of observations" (ibid.). The resulting

data included "questionnaires, interviews, materials analysis, and most importantly, observations of classroom teaching" (ibid., p. 44).

As noted above, both observing and asking are crucial because the washback phenomenon influences both perceptions and behaviors of a wide variety of participants. For this reason, it is essential to investigate both what people do and their interpretations of and reasons for those actions:

> Observations on their own can only reveal part of what is happening within any educational setting: the observers can see what is going on but they may not understand all they see. The other forms of data gathering, though, will be equally uninformative if not accompanied by an analysis of teaching. Without observations the researchers are unlikely to know all the questions they should be asking and may not understand (or be sufficiently critical of) the answers they are given. (Wall & Alderson, 1993, p. 65)

Therefore, triangulation should be incorporated as a methodological cornerstone in any serious investigation of washback.

Quantitative and Qualitative Data. Another issue related to triangulation is the question of whether we should use quantitative data or qualitative data, or both, to investigate washback. The value of using both types of data (a form of methodological triangulation) is explained by Sturman (1996), who investigated students' reactions to registration and placement procedures at two English-language schools in Japan. He used both the students' written open-ended comments (qualitative data) and their responses to a Likert scale (quantitative data) to investigate their perceptions of the oral interview and placement test. Sturman notes,

> The two different types of data give different types of information. Perhaps it would be too easy either to dismiss uncomfortable statistical information, if there wasn't also written evidence of a problem, or to consider written comments to represent the views of just one or two disgruntled students, without further statistical evidence to determine how widespread the views are. (1996, p. 350)

Sturman explains that the value of the open-ended written comments was that they "allow the students' depth of feeling to be expressed," while the advantage of the quantitative data is that they provide "the opportunity to see how representative the written comments are and whether these comments are distributed randomly through the sample" (ibid.). Thus the two different types of data provide a balance of evidence.

Allwright and Bailey (1991) have argued that the debate over whether to use quantitative or qualitative data is not helpful for the research questions posed in today's investigations of language learning and teaching. This is partly because the distinction between quantitative data and qualitative data is somewhat simplistic. In fact, Allwright and Bailey made a case for considering two different dimensions: whether the data are quantitatively or qualitatively *collected*, and whether they are quantitatively or qualitatively *analyzed*. Language testing researchers have typically used quantitatively collected data (such as language learners' test scores and background information) which was then

quantitatively analyzed (for instance, through correlation analyses or tests for statistically significant differences). However, in any investigation of washback we must consider both behaviors and perceptions: What do people do and what do people believe? For this reason, it will be beneficial to utilize both quantitative and qualitative data collection and analysis in future studies of language testing washback.

Additional Research Focus Issues

This literature review has noted many areas where further research on washback is needed. With the advent of the TOEFL 2000 we will have opportunities for extending our investigations of washback to a worldwide database that includes test-takers from many different first languages and cultural backgrounds. The findings of many studies have indicated that language testing washback does influence teaching to some extent (though clearly in content more than in methods). However, we still lack the strong "evidential link," in Messick's terms, that would demonstrate that washback influences language learning, either negatively or positively.

In addition to further research on program washback and much-needed research on learner washback, some other areas that may not be so apparent are also worth investigating in the context of TOEFL 2000. These include seasonality, self-assessment and autonomous learning, and the influence of computer-based testing.

Seasonality. In discussing the relationship between research and what teachers know, Freeman has described a "seasonality" in teaching that appears to be an applicable concept in washback investigations as well. Freeman (1996) notes,

> All teachers learn very early in their careers that teaching and learning have a deeply seasonal rhythm. In North American classrooms, September is different from December, especially just before the holidays, which is different from March, which is different from early June . . . . Although this seasonality is generally trivialized as common sense, it is integral to how teachers plan, how they conduct lessons, and how they manage groups of learners. (pp. 98-99)

This same seasonality has been noted, but not yet specifically investigated, in studies of washback.

For example, Watanabe (1996, p. 331) has suggested that timing of researchers' observations may influence what we discover about washback. He observed Teacher A when the exams were six months away, but Teacher B was observed when the exams were imminent. Watanabe thought that Teacher B's teaching was more closely related to the exam than was Teacher A's. However, this perception was not shared by the teachers in the interview data (although the coordinator thought that both personality variables and the time factor might have influenced the teaching). Watanabe suggests further research to investigate this issue.

Another example of such seasonality is found in Shohamy, Donitsa-Schmidt, and Ferman (1996). Regarding the high-stakes EFL test in Israel, these authors note,

> Teachers report that as the exam date approaches, teaching becomes substantially intensified: "I will spend much more time in oral activities a month or so before the exam" and "I feel that not enough time has been spent on oral activities and I definitely need to work on oral activities more intensely". (p. 308)

Teachers also reported that their students would be excused from classes on the day of the exam and for two days before it (ibid.).

A more extreme case of students being excused from class to prepare for important exams is reported by Wall and Alderson (1993), who state that when observers in their baseline data collection phase went to schools for a third round of observations, they found that "many of the classrooms they had visited had 'dissolved:' schools had stopped giving classes so that students could study on their own" (p. 50). They also report that "most teachers follow the textbook during the first two terms of the year" (ibid., p. 61) but that the third term is very different because

> teachers finish or abandon their textbooks and begin intensive work with past papers and commercial publications to prepare their students for the exam. At this point there is an obvious exam impact on the content of the teaching. (ibid., pp. 61-62)

This kind of test-preparation wave is probably most observable in high-stakes, public examinations that are given regularly but infrequently, such as the yearly administration of the RUE and the HKCEE in Hong Kong, or the annual university entrance examinations in Japan (Buck, 1988; Ingulsrud, 1994).

Whether or not such time-related issues are observable in the case of the TOEFL 2000 remains an open question. Given the likelihood of flexible exam scheduling that may be introduced by computer-based testing, one might expect not to see this sort of dramatic seasonal evidence of washback from the TOEFL 2000. However, the seasonality issue is certainly worth investigating.

Self-Assessment, Autonomous Learners, and Washback. In recent years, concepts such as the "learner-centered curriculum" (Nunan, 1988) and the "task-based syllabus" have greatly altered what we believe about learning and about what teachers and learners should do in classrooms. These curricular and pedagogical innovations have been paralleled and partly driven by second language acquisition research on interlanguage development, input and interaction, learner differences, and cognitive styles. It is beyond the scope of this monograph to review that body of research or the resulting principles and practices of learner-centered teaching that it has influenced. However, we will explore briefly how issues of learner autonomy relate to possible washback research in the context of the TOEFL 2000.

Learner autonomy refers to the philosophy that students should have a large amount to say about what, how, and how fast they learn. The concept incorporates principles of choice, intrinsic motivation, attention focus, and personal evaluation. As Cotterall (1995) explains,

> The main characteristic of autonomy as an approach to learning is that students take some significant responsibility for their own learning over and above responding to instruction. Learners who are autonomous might take responsibility by setting their own

goals, planning practice opportunities, or assessing their progress. (p. 219; see also Nunan, 1996a, p. 15)

The autonomous learner develops his own internal values with regard to judging progress on the material and/or skills to be learned. This ownership and self-regulation are thought to develop greater locus of control and deeper processing of the material at hand (van Lier, personal communication, 1994). Thus, the issue of learner autonomy and responsibility is directly related to Alderson and Wall's (1993) 10th washback hypothesis: "A test will influence the degree and depth of learning" (p. 120).

The autonomous learner does not have to be an isolated individual separate from a program (i.e., here we are not necessarily discussing individual learning or individualized instruction). Rather, we are concerned about the language development of the individual learner, whether he is studying independently, in a self-access center, or in a class. Perhaps it seems strange to focus on washback to the individual learner in the context of the TOEFL test, a standardized norm-referenced test administered to thousands of people annually. Yet it is only in the mind of the individual learner that actual learning, the optimal result of positive washback, can occur.

Relatively little has been written about the impact of an autonomous learning philosophy on assessment, although there is an emerging body of literature related to individualized assessment and self-assessment. The anthology edited by de Jong and Stevenson (1990) provides an interesting collection of articles (and discussants' reactions to them) on individualizing the assessment of language abilities. Dickinson, who has published widely on individualized instruction (see, e.g., Dickinson, 1987), has also written about "self-assessment as an aspect of autonomy" (1982). Brindley's (1989) book about assessing achievement in the learner-centered curriculum has a useful chapter on criterion-referenced assessment, that includes a very helpful section on self-assessment. Other authors have discussed self-assessment relative to placement (Shaw, 1980), self-access learners (Gardner, 1996), classroom teaching and learning (Lewis, 1990), and ESL curricula (Rolfe, 1990). Cohen (1994, pp. 197-206) offers a useful summary of some ideas on self-assessment.

The topic of self-assessment has also received increasing attention from language testing researchers in recent years (see, e.g., LeBlanc & Painchaud, 1985; Oskaarson, 1980; and von Elek, 1985). The following comments from von Elek (1985, p. 60) explain the direct relationship of self-assessment to autonomous learning and positive washback. According to von Elek, self-assessment:

1. enables learners to assume greater responsibility for the assessment of their proficiency and their progress;

2. it enables them to diagnose their weak areas and to get a realistic view of their overall ability and their skills profile;

3. it enables them to see their present proficiency in relation to the level they wish to attain;

4. it helps them to become more motivated and goal oriented.

Thus self-assessment and learner autonomy are linked, because developing internal criteria for success is one of the key characteristics of autonomous learning.

Alderson (1990) ties computer-based testing to learner autonomy. He discusses the kinds of support that software can offer the learners, the computer's capacity for immediate feedback, and the element of choice introduced (e.g., over when to take a computer-delivered test). Alderson suggests that the learner can enter ratings regarding his degree of confidence in completing a task or item and self-ratings of his ability on that particular item or task. The computer can then compare the test taker's ratings of confidence and ability to his actual performance patterns. Alderson states that "such information clearly has implications for learners' self-awareness, and for learner training programmes in which learners' perceptions of their ability can be explored" (ibid., p. 26).

Messick (personal communication) has suggested that a self-assessment questionnaire accompanying the TOEFL 2000 could provide useful data for correlation studies on washback:

> First it would be important, as Alderson and Wall [1993] testify, to obtain baseline data with the current TOEFL test. By compiling the questionnaire results, one can describe the current practices emphasized in the preparation programs taken by the current test-takers as well as their personal efforts to prepare for the TOEFL test. More importantly, one can correlate questionnaire responses and/or scale scores with TOEFL scores and subscores to reveal which program practices and individual learning strategies, if any, are related to TOEFL proficiency outcomes.

Elsewhere, Messick has pointed out that apparent washback needs to be "evidentially linked" to a test in order for washback claims to be convincing.

> More important still, by examining questionnaire results and TOEFL score correlates over time, as TOEFL 2000 is introduced, one can document how the teaching and learning processes change in tandem with the changed TOEFL test. One can also document any changes in the *pattern* of correlations of questionnaire responses with TOEFL scores, which would be a compelling indicator of washback, one way or the other. (Messick, personal communication)

Such data could be collected very economically in a computer-based test (Alderson, 1987), such as the TOEFL 2000.

Computer-Based Testing and Washback. Another area of fruitful investigation related to washback and the TOEFL 2000 stems from its delivery system as a computer-based test and its theoretical measurement basis in Item Response Theory (IRT). The combination of the computer-based delivery system and IRT will allow the TOEFL 2000 to deliver to each test-taker a proficiency exam that is targeted specifically for his or her own level of English development. This capacity overcomes what has been a major drawback of standardized paper-and-pencil tests. Hagiwara (1983) explains the problem with regard to placement tests, but his comments are also applicable to proficiency tests:

> The crux of the matter is that no standardized test appears to have a range of items wide enough to differentiate between [levels] I and II as well as between IV and V. The ideal solutions would call for the construction of such a test; but a project of this scope would entail a considerable amount of time, effort and financial support. (p. 30)

Hagiwara wrote this comment before computer-based testing was a fully viable option. However, new developments in computer-based testing, such as the TOEFL 2000 project, could overcome the problems Hagiwara describes with the range of item difficulty. (It is beyond the scope of this monograph to review computer-based testing in any depth, but helpful sources of information on the use of computers in language testing include Alderson, 1987; Dunkel, 1991; Larsen, 1987; Laurier, 1991; Stansfield, 1986; and Tung, 1985.)

Henning (1987) discusses the advantages of computer-based language testing using Item Response Theory. One of the most notable is the efficient use of test time:

> Once items have been calibrated for difficulty, it is possible to select items to match the known ability ratings of the examinees. Since only those items are used that are necessary to measure the ability of examinees, many redundant or superfluous items can be deleted from the test. The result is a test that can be administered in less time, with less fatigue and boredom for the examinees, and with less expense for the examiners. And this can be accomplished without the sacrifice of test reliability and validity. (p. 110)

This efficiency is due, in part, to the computer's branching capability. Henning explains that computer adaptive testing

> permits the determination of the sequence of items encountered to be based on the ongoing pattern of success and failure experienced by the examinee. Most commonly, such an approach would, for an examinee who experienced success with a given item, result in the purposeful presentation of an item of greater difficulty. The examinee who experienced failure with a given item would next encounter an item of lower difficulty. Some variation of this process would continue in an iterative fashion until it was determined that sufficient information had been gathered about the ability of the examinee to permit termination of the test. (ibid., p. 136)

For these reasons, the TOEFL 2000 should lead to more efficient use of test time (and possibly therefore better attitudes on the part of the test-takers), less measurement error (due to decreased guessing on unnecessarily difficult items), and potentially more positive washback.

Henning also discusses score reporting in the context of computer-based testing and IRT (ibid., pp. 115-116). He gives an example of a diagnostic score report that provides the test-taker with more information than is the case with most widely used proficiency tests. Henning notes that, "Such forms are easily constructed, require a minimal number of pre-calibrated items, and provide a wealth of valuable information to students, teachers and administrators" (ibid., p. 116). Presumably such information

would contribute to positive washback, especially since it can be delivered immediately. Whether or not it does is an empirical question that can be investigated in the context of the TOEFL 2000.

The TOEFL program staff has conducted three studies to investigate the possible impact of the TOEFL 2000's computer-based delivery system on the test-takers. A scale was developed to assess computer familiarity (Eignor, Taylor, Kirsch, & Jamieson, 1997; Kirsch, Jamieson, Taylor, & Eignor, 1997), and a tutorial program was developed to familiarize test-takers with computer-based testing (Taylor, Jamieson, Eignor, & Kirsch, 1997). These authors found no relationship between the test-takers' levels of computer familiarity and their TOEFL scores. However, future research may reveal that computer familiarity is an area where washback could occur. That is, in preparing for the TOEFL 2000, students practicing with computer-based test preparation materials could actually increase their computer familiarity.

Closing Remarks

In 1984, Hale, Stansfield, and Duran published a summary of studies involving the TOEFL test from 1962 to 1983. Although that report included nearly a hundred different research projects, none of them were empirical investigations of washback from the TOEFL test. No follow-up summary of studies involving the TOEFL test has been published in the 1990's (Carol Taylor, personal communication). However, if such a summary were published in 2004, I would hope it would include a variety of research projects investigating language testing washback and the TOEFL 2000.

References

Alderson, J. C. (1987). *Innovation in language testing: Can the micro-computer help? Special report number 1, Language Testing Update.* Lancaster: University of Lancaster.

Alderson, J. C. (1990). Learner-centred testing through computers: Institutional issues in individual assessment. In J. H. A. L de Jong & D. K. Stevenson, (Eds.), *Individualizing the assessment of language abilities* (pp. 20-27). Clevedon: Multilingual Matters.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing 13(3),* 280-297.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics 14(2),* 115-129.

Alderson, J. C., & Wall, D. (1996). Editorial. *Language Testing 13(3),* 239-240.

Allwright, D. (1988). *Observation in the language classroom.* London: Longman.

Allwright, D., & Bailey, K. M. (1991). *Focus on the language classroom: An introduction to classroom research for language teachers.* New York: Cambridge University Press.

American Language Academy. (1997). [Advertising flyer]. *TOEFL Mastery: Computer-based TOEFL test preparation.* Rockville, MD: American Language Academy, ESL Software Department.

Andrews, S. (1994b). The washback effect of examinations: Its impact upon curriculum innovation in English language teaching. *Curriculum Forum (1),* 44-58.

Andrews, S. (1994a). Washback or washout? The relationship between examination reform and curriculum innovation. In D. Nunan, R. Berry, & V. Berry, (Eds.), *Bringing about change in language education: Proceedings of the International Language in Education Conference 1994,* (pp. 67-81). Hong Kong: University of Hong Kong.

Andrews, S., & Fullilove, J. (1994). Assessing spoken English in public examinations – why and how? In J. Boyle & P. Falvey (Eds.), *English language testing in Hong Kong* (pp. 57-86). Hong Kong: Chinese University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

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

Bailey, K. M. (1996a). Working for washback: A review of the washback concept in language testing. *Language Testing 13(3)*, 257-279.

Bailey, K. M. (1996b). The best laid plans: Teachers' in-class decisions to depart from their lesson plans. In K. M. Bailey & D. Nunan (Eds.), *Voices from the language classroom: Qualitative research in second language education* (pp. 15-40). Cambridge: Cambridge University Press.

Berry, V. (1994). Current assessment issues and practices in Hong Kong: A preview. In D. Nunan, R. Berry, & V. Berry, (Eds.), *Bringing about change in language education: Proceedings of the International Language in Education Conference 1994* (pp. 31-34). Hong Kong: University of Hong Kong.

Boyle, J., & Falvey, P. (Eds.). (1994). *English language testing in Hong Kong*. Hong Kong: Chinese University Press.

Brindley, G. (1989). *Assessing achievement in the learner-centred curriculum*. Sydney: Macquarie University, National Centre for English Language Teaching and Research.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Broukal, M., & Nolan-Woods, E. (1995). *NTC's preparation for the TOEFL*. Lincolnwood, IL: National Textbook Company.

Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations. *JALT Journal (10),* 12-42.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics (1),* 1-47.

Chapelle, C., & Douglas, D. (1993). Foundations and directions for a new decade of language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 1-22). Alexandria, VA: TESOL.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education 11(1),* 38-54.

Cheng, L. (forthcoming). *Teacher perspectives and actions toward a public examination change.* Unpublished manuscript. Hong Kong: University of Hong Kong, Department of Curriculum Studies.

Clark, J. L. D. (1983). Language testing: Past and current status—directions for the future. *Modern Language Journal* 67*(4),* 431-443.

Cohen, A. D. (1984). On taking tests: What the students report. *Language Testing 1(1),* 70-81.

Cohen, A. D. (1994). *Assessing language ability in the classroom* (2nd ed.). New York: Heinle and Heinle.

Cotterall, S. (1995). Developing a course strategy for learner autonomy. *ELT Journal 49(3),* 219-227.

de Jong, J. H. A. L., & Stevenson, D. K. (Eds.). (1990). *Individualizing the assessment of language abilities*. Clevedon: Multilingual Matters.

Denzin, N. K. (1970). *Sociological methods: A source book.* Chicago, IL: Aldine.

Dickinson, L. (1982). *Self-assessment as an aspect of autonomy*. Edinburgh: Scottish Centre for Overseas Education.

Dickinson, L. (1987). *Self-instruction in language learning*. Cambridge: Cambridge University Press.

Dunkel, P. (Ed.). (1991). *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House.

Eignor, D., Taylor, C., Kirsch, I., & Jamieson, J. (1998). *Development of a scale for assessing the level of computer familiarity of TOEFL examinees* (TOEFL Research Report No. 60). Princeton, NJ: Educational Testing Service.

Finnochiaro, M., & Sako, S. (1983). *Foreign language testing: A practical approach*. New York: Regents Publishing Company.

Fredericksen, N. (1984). The real test bias: Influences on testing and teaching. *American Psychologist (39),* 193-202.

Freeman, D. (1996). Redefining the relationship between research and what teachers know. In K. M. Bailey & D. Nunan (Eds.), *Voices from the language classroom: Qualitative research in second language education* (pp. 88-115). Cambridge: Cambridge University Press.

Fullilove, J. (1992). The tail that wags. *Institute of Language in Education Journal (9),* 131-147.

G-TELP (General Test of English Proficiency) Information Bulletin. (1990). San Diego: San Diego State University, International Testing Service Center, College of Extended Studies.

G-TELP (General Test of English Proficiency). (undated). Seoul: G-TELP Committee of Korea.

Gardner, D. (1996). Self-assessment for self-access learners. *TESOL Journal 5(3),* 18-23.

Green, D. Z. (1985). Developing measures of communicative proficiency: A test for French immersion students in grades 9 and 10. In P. Hauptman, R. LeBlanc, & M. Wesche (Eds.), *Second language performance testing* (pp. 215-227). Ottawa: University of Ottawa Press.

Haas, G. J. (1990). English language testing: The view from the admissions office. In D. Douglas (Ed.), *English language testing in U.S. colleges and universities*. Washington, DC: National Association for Foreign Student Affairs.

Hagiwara, P. (1983). Student placement in French: Results and implications. *Modern Language Journal 67(1),* 23-32.

Hale, G. A., Stansfield, C. W., & Duran, R. P. (1984). *Summaries of studies involving the Test of English as a Foreign Language, 1963-1982* (TOEFL Research Report #16). Princeton, NJ: Educational Testing Service.

Hammersley, M., & Atkinson, P. (1983). *Ethnography: Principles and practice*. London: Routledge.

Hamp-Lyons, L. (1996). *Ethical test preparation: The case of TOEFL*. Paper presented at the Language Testing Research Colloquium, Chicago, IL.

Hart, D., Lapkin, S., & Swain, M. (1987). Communicative language tests: Perks and perils. *Evaluation and Research in Education 1(2),* 83-93.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. New York: Newbury House.

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practices (12),* 20-25.

Heyneman, S. P. (1987). Uses of examinations in developing countries: Selection, research, and education sector management. *International Journal of Educational Development 7(4),* 251-263.

Hilke, R., & Wadden, P. (1996). The TOEFL and its imitators: Analyzing the TOEFL and evaluating TOEFL-prep texts. *PASAA: A Journal of Language Teaching and Learning in Thailand (26),* 53-67.

Huberman, M., & Miles, M. (1984). *Innovation up close*. New York: Plenum.

Hughes, A. (1988). Introducing a needs-based test of English language proficiency into an English-medium university in Turkey. In A. Hughes (Ed.), *Testing English for university study* (pp. 134-146). (ELT Documents #127). London: Modern English Publications in association with the British Council.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hughes, A. (1993). *Backwash and TOEFL 2000*. Unpublished manuscript, University of Reading.

Ingulsrud, J. E. (1994). An entrance test to Japanese universities: Social and historical contexts. In C. Hill & K. Parry (Eds.), *From testing to assessment: English as an international language* (pp. 61-81). London: Longman.

Johnson, F., & Wong, C. L. K. L. (1981). The interdependence of teaching, testing and instructional materials. In J. A. S. Read (Ed.), *Directions in language testing* (pp. 277-302). Singapore: Regional Language Centre.

Johnson, K. (1992). The instructional decisions of pre-service ESL teachers: New directions for teacher preparation programs. In J. Flowerdew, M. Brock, & S. Hsia (Eds.), *Perspectives on second language teacher education*. Hong Kong: City Polytechnic of Hong Kong.

Kennedy, C. (1988). Evaluation of the management of change in ELT projects. *Applied Linguistics 9(4),* 329-342.

Khaniya, T. R. (1990). *Examinations as instruments for educational change: Investigating the washback effect of the Nepalese English exams*. Unpublished doctoral dissertation, University of Edinburgh.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees* (TOEFL Research Report No. 59). Princeton, NJ: Educational Testing Service.

Lam, H. P. (1994). Methodology washback – an insider's view. In D. Nunan, R. Berry, & V. Berry (Eds.), *Bringing about change in language education: Proceedings of the International Language in Education Conference 1994* (83-102). Hong Kong: University of Hong Kong.

Larsen, J. W. (1987). Computerized adaptive testing. In K. M. Bailey, T. L. Dale, & R. T. Clifford (Eds.), *Language testing research: Selected papers from the 1986 Colloquium* (pp. 1-10). Monterey: Defense Language Institute.

Laurier, M. (1991). What we can do with computerized adaptive testing...and what we cannot do! In S. Anivan (Ed.), *Current developments in language testing* (pp. 244-255). Anthology Series 25. Singapore: Regional Language Centre.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly 19(4),* 673-687.

Lewis, J. (1990). Self-assessment in the classroom: A case study. In G. Brindley (Ed.), *The second language curriculum in action* (pp. 187-213). Sydney: Macquarie University, National Centre for English Language Teaching and Research.

Lougheed, L. (1992). *TOEFL practice book* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall Regents.

Lowe, P. (1988). Introduction. In P. Lowe & C. W. Stansfield (Eds.), *Second language proficiency assessment: Current issues* (pp. 1-10). Englewood Cliffs, NJ: Prentice Hall Regents.

Madaus, G. (1990). *Testing as social technology*. Paper presented at the inaugural annual Boisi Lecture in Education and Public Policy. Boston College, 6 December.

Markee, N. (1993). The diffusion of innovation in language teaching. *Annual Review of Applied Linguistics (13),* 229-243.

Mehrens, W. A., & Kaminsky, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless or fraudulent? *Educational Measurement: Issues and Practices 8(1),* 14-22.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher (1)23,* 13-23.

Messick, S. (1996). Validity and washback in language testing. *Language Testing 13(3),* 241-256.

Morrow, K. (1991). Evaluating communicative tests. In S. Anivan (Ed.), *Current developments in language testing* (pp. 111-118). Anthology Series 25. Singapore: Regional Language Centre.

Nunan, D. (1988). *The learner-centred curriculum: A study in second language teaching*. Cambridge: Cambridge University Press.

Nunan, D. (1996a). Towards autonomous learning: Some theoretical, empirical and practical issues. In R. Pemberton, E. S. S. Li, W. Or, & H. Pierson (Eds.), *Taking control: Autonomy in language learning* (pp. 13-26). Hong Kong: Hong Kong University Press.

Nunan, D. (1996b). Hidden voices: Insiders' perspectives on classroom interaction. In K. M. Bailey & D. Nunan (Eds.), *Voices from the language classroom: Qualitative research in second language education* (pp. 41-56). Cambridge: Cambridge University Press.

Oller, J. W., Jr. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Oskaarson, M. (1980). *Approaches to self-assessment in foreign language learning*. Oxford: Pergamon Press.

Pearson, I. (1988). Tests as levers of change (or 'putting first things first'). In D. Chamberlain & R. Baumgartner (Eds.), *ESP in the classroom: Practice and evaluation* (pp. 98-107). *ELT Documents #128*. London: Modern English Publications in association with the British Council.

Pierce, B. N. (1992). Demystifying the TOEFL reading test. *TESOL Quarterly 26(4),* 665-691.

Popham, W. J. (1991). Appropriateness of teachers' test preparation practices. *Educational Measurement: Issues and Practices 10(1),* 12-15.

Raimes, A. (1990). The TOEFL Test of Written English: Causes for concern. *TESOL Quarterly 24(3),* 427-442.

Rogers, B. (1993). *The complete guide to the TOEFL.* Boston: Heinle and Heinle.

Rolfe, T. (1990). Self- and peer-assessment in the ESL curriculum. In G. Brindley (Ed.), *The second language curriculum in action* (pp. 163-186). Sydney: Macquarie University, National Centre for English Language Teaching and Research.

Schutz, N. W., Jr., Derwing, B., Palmer, I., & Steed, J. F. (1991). *Breaking the TOEFL barrier.* Englewood Cliffs, NJ: Prentice Hall.

Sharpe, P. J. (1996). *Barron's students' #1 choice: How to prepare for the TOEFL* (eighth edition). Woodbury, NY: Barron's.

Shaw, P. (1980). Comments on the concept and implementation of self-placement. *TESOL Quarterly 14(2),* 261-262.

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal 76(4),* 513-521.

Shohamy, E. (1993a). The power of tests: The impact of language tests on teaching and learning. *NFLC Occasional Paper.* Washington, DC: National Foreign Language Center.

Shohamy, E. (1993b). A collaborative/diagnostic feedback model for testing foreign languages. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 185-202). Alexandria, VA: TESOL Publications.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing 13(3),* 298-317.

Shohamy, E., Reves, T., & Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *ELT Journal 40(3),* 212-220.

Smallwood, I. M. (1994). Oral assessment: A case for continuous assessment at HKCEE level. *New Horizons: Journal of Education, Hong Kong Teachers' Association 35,* 68-73.

Spolsky, B. (1990). Social aspects of individual assessment. In J. H. A. L. de Jong & D. K. Stevenson (Eds.), *Individualizing the assessment of language abilities* (pp. 3-15). Clevedon: Multilingual Matters.

Spolsky, B. (1994). The examination-classroom backwash cycle: Some historical cases. In D. Nunan, R. Berry, & V. Berry (Eds.), *Bringing about change in language education: Proceedings of the International Language in Education Conference 1994* (pp. 55-66). Hong Kong: University of Hong Kong.

Stansfield, C. W. (Ed.). (1986). *Technology and language testing*. Washington, DC: TESOL Publications.

Stoller, F. (1994). The diffusion of innovations in intensive ESL programs. *Applied Linguistics 15(3),* 300-327.

Sturman, P. (1996). Registration and placement: Learner response. In K. M. Bailey & D. Nunan (Eds.), *Voices from the language classroom: Qualitative research in second language education* (pp. 338-355). Cambridge: Cambridge University Press.

Swain, M. (1984). Large-scale communicative testing: A case study. In S. L. Savignon & M. Berns (Eds.), *Initiatives in communicative language teaching* (pp. 185-201). Reading, MA: Addison Wesley.

Swain, M. (1985). Large-scale communicative testing: A case study. In Y. P. Lee, A. C. Y. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 35-46). Oxford: Pergamon Press.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (TOEFL Research Report No. 61). Princeton, NJ: Educational Testing Service.

Tung, P. (1985). Computerized adaptive testing: Implications for language test developers. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 12-27). Ottawa: University of Ottawa Press.

van Lier, L. (1988). *The classroom and the language learner: Ethnography and second-language classroom research*. London: Longman.

Von Elek, T. (1985). A test of Swedish as a second language: An experiment in self-assessment. In Y. P. Lee, A. C. Y. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 47-55). Oxford: Pergamon Press.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing 13(3),* 334-354.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing 10(1),* 41-69.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing 13(3),* 318-333.

Wesche, M. B. (1983). Communicative language testing in a second language. *Modern Language Journal 67(1),* 41-55.

Wesche, M. B. (1987). Second language performance testing: The Ontario Test of ESL as an example. *Language Testing 4(1),* 28-47.

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

Test of English as a Foreign Language
P.O. Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 609-771-7100
E-mail: toefl@ets.org
Web site: http://www.toefl.org

58713-16347 • Y79M.75 • 253709 • Printed in U.S.A.