Aiming for positive washback: a case study of international teaching assistants

Shahrzad Saif *Université Laval*

The aim of this study is to explore the possibility of creating positive washback by focusing on factors in the background of the test development process and anticipating the conditions most likely to lead to positive washback. The article reports on a multiphase empirical study investigating the washback effects of a needs-based test of spoken language proficiency on the content, teaching, classroom activities and learning outcomes of the ITA (international teaching assistants) training program linked to it. As such, the conceptual framework underlying the study differs from previous models in that it includes the processes before test development and test design as two main components of washback investigation. The analysis of the data – collected from different stakeholders through interviews, observations and test administration at different intervals before, during and after the training program – suggests a positive relationship between the test and the immediate teaching and learning outcomes. There is, however, no evidence linking the test to the policy or educational changes at an institutional level.

I Introduction

This article discusses an institution-wide attempt to create positive washback through the introduction of a test of spoken language ability specifically designed for international teaching assistants (ITAs) functioning in an English-speaking university, i.e. the University of Victoria (UVic), Canada. The main goal of this study is to examine how the high-stakes performance test developed based on the practical needs of ITAs and the educational context in question will influence teaching activities and learning outcomes of an ITA program linked to it.

The term 'washback' is, therefore, used in this study to refer to the effects of tests on course content, teaching, learning, and classroom

Address for correspondence: Shahrzad Saif, Assistant Professor, Département de langues, linguistique et traduction, Université Laval, Québec, Québec G1K 7P4, Canada; email: shahrzad.saif@lli.ulaval.ca

Language Testing 2006 23 (1) 1–34 10.1191/0265532206lt322oa

activities. Also, in its attempt to create positive washback, the study focuses on the needs of ITAs and the educational context as perceived by various stakeholders. This includes gathering information about the characteristics of ITAs, the educational context in which they function, the language tasks they engage in, and baseline information about past ITA testing and teaching practices at UVic. Baseline information has been sought using two main sources of information: examination of available documents on the testing instrument and ITA course materials, and open-ended interviews with people involved in ITA programs in the past (i.e. English Language Centre teachers and administrators). Because ITA courses had not been offered for some time before this study started, it was not possible to collect quantitative baseline data with respect to ITAs' language learning outcomes in the past. The effect of the test developed in this study on the students' learning is, therefore, examined on its own terms.

As a result of internationalization as one of the goals of UVic's strategic plan, when this study started in late 1996, increasing numbers of foreign graduate students were functioning as teaching, laboratory, and research assistants at the university. This highlighted the need for the establishment of a testing mechanism to determine whether ITAs had a sufficient level of spoken English proficiency to be able to communicate successfully with their undergraduate students in instructional settings. There was equally a need for a training program into which ITAs with unsatisfactory test results could be channeled. However, as the new test was seen by the school administrators as part of an attempt to improve the oral communication abilities of ITAs, it was expected to positively influence the objectives and activities of the training program linked to it. In other words, the development of the new test was viewed not as an end to itself but as a potential means of achieving classroom success.

II Test washback: theoretical background

Research into the effects of tests on teaching and learning activities, referred to as 'washback' or 'backwash' in the applied linguistics literature (Hughes, 1989; Khaniya, 1990; Alderson, 1991; Alderson and Wall, 1993; among others), dates back at least four decades. However, a glance at the literature on language teaching and testing reveals that there is considerable variation in the way different

authors have theoretically portrayed this phenomenon. While some authors consider tests as having nothing but negative consequences for teaching methodology and syllabus content (Vernon, 1956; Wiseman, 1961), others look at tests more positively with important implications for curriculum and as potential instruments for educational reform (Swain, 1985; Alderson, 1986; Pearson, 1988; Hughes, 1989; among others). There have also been views with respect to the significance of a positive washback effect for test validity (Morrow, 1986; Frederiksen and Collins, 1989). Messick (1996), however, convincingly argues that factors other than the test itself might prevent the intended positive washback from happening, and that a lack of desired washback effect does not by itself render a test as invalid unless there is sufficient evidence that the negative washback effect is a direct result of the test and its lack of construct validity. He further suggests that to accomplish positive washback, 'rather than seeking washback as a sign of test validity, seek validity by design as a likely basis for washback' (p. 252) by enhancing test tasks and content so that they adequately represent test constructs. Similarly, Turner (2001: 139), based on her experience with empirically derived rating scales in high-stakes performance testing, identifies the need 'to consider the potential of washback effects at different times throughout the testing cycle (i.e. anywhere decisions need to be made concerning evaluation – needs analysis, purpose of test … etc.)'. An important implication of this is that intentional moves towards positive washback, like the one discussed in this study, should include test design as the means of achieving washback. Efforts to achieve positive washback can be, therefore, enhanced by conducting washback investigations within the confines of a conceptual framework that includes the test development process as part of the investigation process; one that, besides specifying the areas most likely to be affected by the test, allows the researcher to incorporate learners' language use needs, and control for the institutional stakeholders' goals and strategies from the outset.

This is by no means the first time that the need for a theoretical framework has been pointed out (see Bailey, 1996; Wall, 2000). It should be noted, however, that previous models (Hughes, 1993; Bailey, 1996) do not include test design based on the needs of the learners in individual contexts as part of the washback investigation process. They discuss washback issues 'specifically in the context of external-to-program tests' (Bailey, 1996: 257) for learners aiming for a given 'general proficiency' level. As such, these models may be appropriately adapted only to studies that seek to identify possible

washback effects of already existing tests that may or may not be valid for the uses to which they are put. In fact, empirical research (Wall and Alderson, 1993; Shohamy *et al.*, 1996; Alderson and Hamp-Lyons, 1996; Watanabe, 1996; Cheng, 1997; Andrews *et al.*, 2002; among others) has overwhelmingly focused on the effects of such tests on different aspects of corresponding curricula and produced valuable insights into the effects of – mostly high-stakes – tests on such areas as materials development, teaching activities, course content, teachers' methodology and learning. Wall (2000), nevertheless, underlines the need for empirical evidence – with respect to the washback effects of new tests on students' learning – by using 'an independent test which measures the "right" things (the aims of the curriculum)' (p. 502). Technically speaking, this too implies the inclusion of the 'test design' as a component of washback investigation.

So, given (1) the setting of this study, which required the introduction of a new test as part of an intentional attempt to positively influence the content, teaching activities and learning outcomes of an ITA training program; and (2) the 'specific purposes' (rather than 'general proficiency') nature of ITA learning, the conceptual framework presented below is introduced to guide practical research in this study. The components of the model systematically represent the study's major focus areas grouped under three categories of 'needs', 'means', and 'consequences'. The model illustrates two major lines of connection to be pursued in this study with respect to the test: first, the needs and objectives of the population and the educational context in question, which directly or indirectly affect the type, purpose, and content of the test, its development and implementation; and, second, the potential effects of test use on classroom teaching and learning activities. See Figure 1.

The incorporation of the needs and objectives of the educational setting and the test development process as a point of departure for this investigation would then exclude the test as the sole initiator of washback operation (as implied by previous models). Instead, it would allow the inclusion of certain areas of potential impact on the participants of this particular context thereby facilitating the 'washback to the program' (Bailey, 1996). For example, the model suggests that the test be developed with respect to a theoretical framework in conformity with the test objectives so that the same theoretical line of thought can be followed (for example, by teachers and material developers) in all future decisions made with respect to material development and teaching methodology. Moreover, to

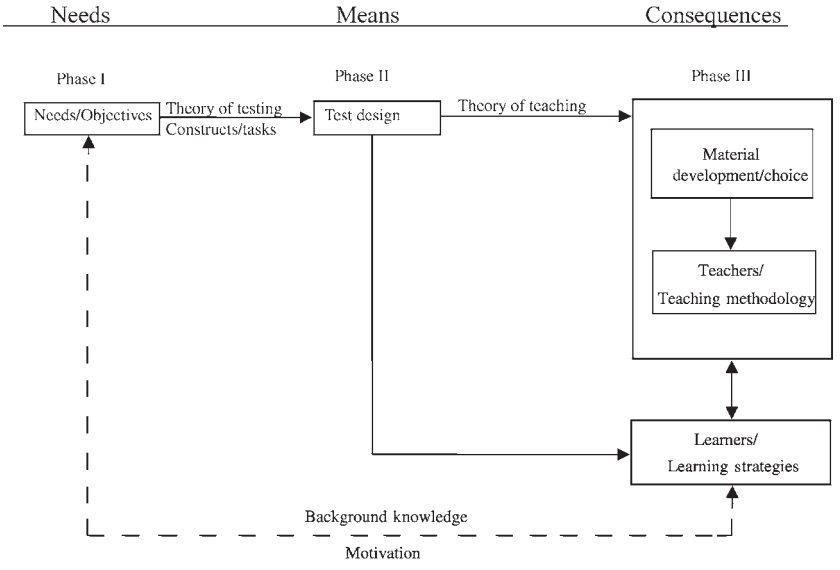

Figure 1 A conceptual framework for washback

enhance desirable learning effects, the model suggests that such factors as learners' motivation and background knowledge – ITAs' previous experience with the target language as well as their topical knowledge – be taken into consideration in the development of the test. The two-way relationship between the components of the model further allows for what Shohamy (1992: 515) calls the involvement of 'the ones expected to carry out change' (in this case, the teacher) in the test development and/or administration process.

Empirical research – with the purpose of examining the possibility of creating washback through the introduction of a new test based on the specific needs of the ITAs – was then carried out in three different phases each of which corresponded to one of the different levels of the model described above.

III Phase I: needs analysis

The above objective necessarily entailed a clear description of the language needs of the ITAs and the educational context in question. As a result, a needs analysis was conducted during the first phase of the study using both inductive and deductive (Berwick, 1989;

Brindley, 1989) approaches. Three major groups of stakeholders – graduate advisors and administrators, undergraduate students, and ITAs – were consulted. Information with respect to the ITAs' subjective needs (e.g. academic roles, wants and expectations, learning needs) was collected through open-ended interviews with stakeholders in various positions. Objective information (e.g. ITAs' personal data, present language proficiency, language use contexts), on the other hand, was gathered through a questionnaire based on Munby's 1981 framework. Information about the university's past ITA teaching and testing practices was gathered separately by examining available documents and interviewing the administrators and ESL teachers in the English Language Centre (ELC).

1 Graduate advisors and administrators

In a general assembly of Faculty of Graduate Studies (FGS), graduate advisors of different departments were consulted with respect to different issues involving the language proficiency level of ITAs in their department. Of 27 graduate advisors originally invited, 19 attended the meeting. Also in attendance was the director of the graduate records office. Table 1 details graduate advisors' responses to the researcher's questions.

The information from the records office revealed that the university required a minimum TOEFL score of 550/213 for all foreign applicants to the graduate programs even though some departments had set higher TOEFL standards. There was no legislation or policy in effect regarding ITAs' oral and/or written language proficiency. Graduate advisors also indicated that other than the TOEFL score the departments had no language requirements for TA (teaching assistant) assignments. TA positions were assigned by the departments solely based on the departmental educational needs and candidates' academic preparedness for the assigned tasks. Teaching experience, on the other hand, was not a determining factor in TA assignments. Therefore, based on the departments' selection process, all graduate students (native-speaking and ITAs alike) who applied for TA positions needed to possess the required topical knowledge but not necessarily any teaching expertise.

The results of this initial survey also revealed that TAs' jobs at UVic consisted mainly of assisting professors in grading exams, preparing course materials, supervising lab sessions, holding office hours, tutoring students, substituting for professors at times, and

Table 1 Summary of the information provided by graduate advisors

Department	Contexts in which ITAs have to use English				Tasks ITAs have to perform in such contexts						Type/level of language ability required in these contexts		Language ability/ abilities most fundamental to ITAs' dealings with undergraduate students				Number of ITAs who are seriously deficient in these areas		Requirement for hiring ITAs in each department			Main source of ITAs' communication problems in your department			Does the department support the adoption of a screening test for ITAs?	
													Spoken		Written										Yes	No
	Classroom	Office	Laboratory	Other	Preparing materials	Teaching	Supervising lab sessions	Holding tutorials office hours	Marking	Other	Spoken (advanced/ intermediate/low)	Written (advanced/ intermediate/low)	Speaking	Listening comprehension	Writing	Reading comprehension	More than 50%	Less than 50%	TOEFL	Knowledge of the course content	Teaching experience	Language problem	Lack of competency in the content area	Other		
Biochemistry	✓	✓	✓		✓	✓	✓	✓	✓		A	I	✓	✓			✓		✓	✓		✓			✓	
Biology	✓	✓	✓		✓	✓	✓	✓	✓		A	I	✓	✓			✓		✓	✓		✓			✓	
Chemistry		✓	✓		✓		✓	✓	✓		A	I	✓	✓			✓		✓	✓		✓			✓	
Computer science	✓	✓	✓		✓	✓	✓	✓	✓		A	I	✓	✓			✓		✓	✓		✓			✓	
Curriculum		✓			✓			✓	✓		A	I	✓	✓		✓		✓	✓	✓		✓			✓	
Earth sciences	✓	✓	✓		✓	✓	✓	✓	✓		A	I	✓	✓			✓		✓	✓		✓			✓	
Education	✓	✓			✓	✓		✓	✓		A	I	✓	✓		✓		✓	✓	✓		✓			✓	
Electrical engineering	✓	✓	✓		✓	✓	✓	✓	✓		A	I	✓	✓			✓		✓	✓		✓			✓	

(continued)

Table 1 continued

Department	Contexts in which ITAs have to use English				Tasks ITAs have to perform in such contexts						Type/level of language ability required in these contexts		Language ability/ abilities most fundamental to ITAs' dealings with undergraduate students				Number of ITAs who are seriously deficient in these areas		Requirement for hiring ITAs in each department					Main source of ITAs' communication problems in your department				Does the department support the adoption of a screening test for ITAs?	
													Spoken		Written														Yes	No
	Classroom	Office	Laboratory	Other	Preparing materials	Teaching	Supervising lab sessions	Holding tutorials office hours	Marking	Other	Written (advanced/ intermediate/low)	Spoken (advanced/ intermediate/low)	Speaking	Listening comprehension	Writing	Reading comprehension	Less than 50%	More than 50%	Teaching experience	Knowledge of the course content	TOEFL	Other	Lack of competency in the content area	Language problem						
English		✓			✓			✓	✓		I	A	✓	✓		✓	✓			✓	✓							✓	
History		✓			✓			✓	✓		I	A	✓	✓		✓	✓			✓	✓							✓	
Linguistics		✓	✓		✓		✓	✓	✓		I	A	✓	✓		✓	✓			✓	✓			✓				✓	
Mathematics	✓	✓			✓	✓		✓	✓		I	A	✓	✓				✓		✓	✓			✓				✓	
Mechanical engineering	✓	✓	✓		✓	✓	✓	✓	✓		I	A	✓	✓				✓		✓	✓			✓				✓	
Philosophy		✓			✓			✓	✓		A	A	✓	✓		✓	✓			✓	✓							✓	
Physics	✓	✓	✓		✓	✓	✓	✓	✓		I	A	✓	✓				✓		✓	✓			✓				✓	
Political science		✓			✓			✓	✓		I	A	✓	✓				✓		✓	✓			✓				✓	
Psychology	✓	✓			✓	✓		✓	✓		I	A	✓	✓		✓		✓		✓	✓			✓				✓	
Sociology	✓	✓			✓	✓		✓	✓		I	A	✓	✓		✓	✓			✓	✓							✓	
Visual arts		✓			✓			✓	✓		I	A	✓	✓				✓		✓	✓			✓				✓	

teaching independent courses (especially during summer sessions). Therefore, TAs had to display their academic as well as their language skills in such instructional settings as classrooms, laboratories, and during office hours. All graduate advisors believed that functioning in these contexts required the ability to communicate orally at an advanced level. As for ITAs' spoken language proficiency negatively affecting their performance in such contexts, there were different views. Whereas all graduate advisors supported the decision of FGS in adopting a screening mechanism for determining ITAs' language skills, not all of them considered the problem a major one. This was mainly due to the fact that in some disciplines – such as those in Humanities – the ratio of foreign graduate students to domestic ones was lower and TAs in those departments were not assigned independent teaching and lab supervision tasks. Graduate advisors in Sciences and Engineering, on the other hand, felt that the problem was a serious one and were interested in improving ITAs' spoken language ability. This was later justified by the fact that 87% of the students who participated in the experiment were in Engineering and Sciences while only 13% were in Humanities and Fine Arts.

2 Undergraduate students

In a separate survey, 255 randomly selected undergraduate students were interviewed for what they generally thought about international graduate students' performance as TAs. All subjects had either taken a course taught by an ITA or participated in a course whose laboratory and/or office hour session had been run by an ITA. They were specifically asked to comment on ITAs knowledge of the subject matter and language abilities. They were also asked if they thought cultural differences affected in any way the communication taking place between them and ITAs in the above contexts.

A majority of respondents (72%) mentioned they had experienced problems communicating with ITAs. They all characterized ITAs' spoken language problems as the source of the problem. The remaining 28% referred to ITAs' English language proficiency, but did not think it was a major problem interfering with their understanding of the subject matter. Some students in this group even mentioned that they had developed strategies to get around the problem, for example, by trying to find the pronunciation patterns of some ITAs, or asking for clarification. This very point, the lack of the adequate elaboration of the teaching material, was what many respondents in the majority group referred to as a problem caused by ITAs' language

problems. There was, however, no reference by undergraduates to ITAs' academic incompetence as the source of communication problems. Neither did any respondent relate classroom communication problems to cultural differences.

3 International teaching assistants

An adaptation of Munby's 1981 framework (Appendix 1) was used for gathering objective information regarding ITAs' general background and language needs. The questionnaire was completed from data based on interviews held with the participants. In addition, the candidates were asked open-ended questions about what they thought about their spoken English language proficiency level, their knowledge of the subjects they were assigned to, their motivation for accepting TA jobs, their interest in improving their spoken English, and their willingness to take a language course if necessary.

Forty-seven foreign graduate students from different departments participated in this survey. They were from 11 different nationalities with 45% female students and a mean age of 33. Eighty-seven percent of the respondents came from different areas of engineering and sciences. Their average TOEFL score was 597; however, they showed no evidence of their spoken language proficiency level. The respondents indicated they had at least 5 years of formal EFL training with 30% having at least 3 years of teaching experience back home. As for the context of language use, ITAs too referred to classrooms, laboratory, tutorial, and office hour sessions as contexts in which they had to regularly use English. While all respondents considered these settings as professionally and linguistically demanding, only 32% thought they were culturally unfamiliar and formal. All ITAs considered the production and comprehension of the spoken language as the most important language ability needed to communicate with undergraduate students in such contexts.

In answer to the question regarding their academic preparedness for the courses assigned to them in their respective departments, all ITAs echoed their graduate advisors by rating themselves favorably. In contrast, 66% of ITAs expressed, one way or another, concerns about their linguistic abilities and interest in improving their communication skills through an ITA training course. Half of the respondents, however, believed that the course should not be mandatory or a requirement for TA assignments. Eighty-five percent of the ITAs considered teaching assistantship as their main source of income, and financial reasons as their primary motivation for taking TA

positions. This was despite the fact that all respondents directly and/or indirectly acknowledged teaching assistantship to be a valuable experience and an avenue for gaining teaching experience at the college level.

4 Baseline study

In the final stage of Phase I, information was gathered with respect to past ITA teaching and testing practices at the university. As already mentioned in Section I, ITA courses had not been offered for a while at UVic, so ELC could not provide any quantitative data with respect to ITAs' learning outcomes in the past. However, since the new test was to be developed as part of an attempt to create washback in a deliberate context, it was important to investigate what teachers thought of the old test and whether or not they believed it had affected the content of their syllabi, classroom activities, and teaching methods. The information was collected through examining the existing documents and interviewing administrators and teaching staff at ELC, the unit responsible for the direction, coordination, teaching, and testing of ESL courses. The results of this investigation pointed to two different types of TA training programs offered at UVic: a week-long orientation program sponsored by FGS, and an ITA course offered by ELC.

A TA orientation program was offered on a regular basis at the beginning of each academic year and provided practical advice and training on a range of teaching-related topics from 'grading' to 'lab instruction' to 'being a TA when English is not your first language' in seminars and workshops run by graduate advisors and professors. On the other hand, the ITA course, referred to as Graduate English Language Program, had been offered sporadically in the past for those graduate students who wished to improve their English language skills, or had been referred to the ELC by their departments. According to the ELC's program director, the students were assigned to the course based on the results of the Speaking Test, an in-house test of speaking that served as both an entrance and exit test for the ITA course. The test was made up of three sections: interview, narration, and discussion. In the interview section, the students were asked to answer questions like 'How long have you been here?', 'What are your future plans?', or 'How did you feel on the day you left your country?', while in the narration section, they were required to tell a story based on a series of connected pictures. The discussion part engaged pairs of students in a 3-minute discussion of a topic

such as 'Do you think there is too much violence on TV?' or 'Do you support capital punishment?' The students were then rated on a 0–6 scale for pronunciation, grammar, vocabulary, and fluency. While grammar, fluency, and vocabulary were each accorded 30% towards the final total, pronunciation was given 10%. The test kit did not provide any information regarding the rationale underlying the choice of the test tasks and topics of discussion except that they allowed a broad sampling of language functions and styles. In short, the test, as it stood, concentrated primarily on topics of general knowledge, non-technical vocabulary and short informal conversation practice. These ability areas, while useful for everyday communication, would not address the communicative needs of ITAs.

An examination of the teachers' syllabi further revealed no clear relationship between the test and teaching activities, nor did the course materials (chosen by teachers) promote the objectives of the test. The lack of uniform teaching materials and a valid testing device had led the teachers to base the content of the course on what they 'felt' was the course objective, which was not necessarily in conformity with ITAs' real language needs. One ESL teacher – who later participated in the administration of the new test (see Section V) – for example, commented that the test was not valid for use in the ITA course and that she had based her teaching on improving the 'teaching abilities' of TAs rather than language abilities promoted by the test. For the departments, however, improving teaching skills was not a priority. In fact, native-speaking TAs were equally inexperienced in teaching, yet they were never required to take a course in ELC. The analysis of another teacher's syllabus revealed that she had understandably eliminated the test from her syllabus and chosen to assess the students on the basis of their performance during the course. The course content, however, included a variety of activities ranging from handout preparation and class presentations on topics such as 'the six most important discoveries in the world' to writing technical papers and CV preparation. A third teacher had decided to work solely on students' common pronunciation problems through laboratory practice at the expense of interactive activities.

IV Phase II: the new test

In this phase of the study, the theoretical model of language testing proposed by Bachman and Palmer (1996) was used as a guiding

rationale for test development. The details of the test development process, together with its reliability and practicality studies have been reported in a previous article (Saif, 2002); however, a summary of the major steps taken in this phase are presented below.

1 TLU tasks and test constructs

As a first step, based on the information gathered from graduate advisors, 'teaching undergraduate courses', 'conducting lab sessions' and 'holding tutorials/office hours' were identified as target language use (TLU) tasks ITAs performed in instructional contexts at UVic. So, to provide a basis for the development of test tasks that corresponded closely to real-life language use tasks, the model of task characteristics (Bachman and Palmer, 1996) was used to specify the characteristics of each TLU task. At the same time, the constructs to be measured by the test were defined in terms of relevant categories of the model of language ability (Bachman and Palmer, 1996) and with direct reference to the objectives of the test, needs of the test-takers, and the characteristics of the context (Table 2).

A few factors were taken into consideration in the definition of the constructs. Because of the interactional nature of the instructional settings and ITAs' need to demonstrate their ability to set and accomplish communicative goals, strategic knowledge was included in the constructs. Topical knowledge, on the other hand, was not included in the constructs because ITAs belonged to different academic disciplines, and 'academic preparedness' was a hiring criterion already ensured by the individual departments to which ITAs belonged. Finally, while the construct definition included those teaching-related language abilities (such as strategic and textual knowledge) commonly used in instructional contexts, 'teaching skills' was not included in the constructs mainly because teaching ability was not a requirement for hiring – native or non-native speaking – TAs at UVic. There were native-speaking TAs who were inexperienced in teaching but never tested for their teaching skills. At the same time, including teaching skills in the constructs could help ITAs with teaching background to invalidly boost their scores by demonstrating teaching strategies that were not necessarily language-related. Besides, both graduate advisors and undergraduate students identified ITAs' 'language' not 'teaching' abilities as the main source of their communication problems. The distinction McNamara (1996) makes between a 'strong' and 'weak' performance hypothesis is,

Table 2 Test constructs

Language knowledge	
• Grammatical knowledge	Ability to draw upon syntactic, lexical and phonological knowledge in production of well-formed, comprehensible utterances: knowledge of grammatical structures, accurate use of them for the purpose of communication; knowledge of general and specialized vocabulary; knowledge of phonological rules.
• Textual knowledge	Ability to organize utterances to form an oral text: knowledge of cohesive devices used to mark the relationships; knowledge of common methods for organizing thoughts.
• Functional knowledge	Ability to create and interpret spoken language in relation to different functions common to instructional settings: how to use language for expressing information, ideas and knowledge (descriptions, classifications, explanations), making suggestions and comments, establishing relationships, and transmitting knowledge.
• Sociolinguistic knowledge	Ability to relate utterances to the characteristics of the setting: use of the standard dialect, relatively formal register.
Strategic competence	Ability to set goals for the communication of the intended meanings, assess alternative linguistic means (especially when there is a linguistic problem preventing the speaker/hearer from completing a default task), and to draw upon the areas of language knowledge for the successful implementation and completion of a chosen task.

therefore, relevant in the context of this study. Test scores would be used to make inferences about the test-takers' ability to use language in instructional settings (weak version), not their ability to perform the TA job in English (strong version).

2 Test tasks

In the final step before developing the test task, the characteristics of TLU tasks mentioned above were reviewed in the light of the test constructs. While the three tasks had a lot of common characteristics, the 'teaching' task which was longer, more speeded, and more textually complex than others was selected as a basis for specifying the characteristics of the test task (Table 3).

Table 3 Characteristics of the test task

Characteristics of the setting	
• Physical characteristics	Location: on-campus classroom, well-lit, comfortable temperature Noise level: normal Materials and equipment: books, notes, black-board, overhead projectors, video-camera, etc. Degree of familiarity: everything familiar to the test-taker except for the video-camera
• Participants	Two ESL instructors from English Language Centre (raters), the researcher, 3 undergraduate students from test-taker's department (raters)
• Time of task	First week of September, weekday afternoons
Characteristics of the input	
Format:	
• Channel	Oral
• Form	Language
• Language	English
• Length	Moderate
• Type	Simple short prompt providing necessary instructions, complex prompts in the form of a question providing context for the speaking task
• Speededness	Unspeeded
• Vehicle	Live
Language of input:	
• Grammatical	General and technical vocabulary, varied grammatical structures
• Textual	Utterances within each prompt properly linked and organized
• Functional	Ideational, manipulative
• Sociolinguistic	Standard dialect, mostly formal register, no cultural references
• Topical characteristics	Same as the one picked by the test-taker
Characteristics of the expected response	
Format:	
• Channel	Oral
• Form	Language and non-language (depending on the subject)
• Language	English
• Length	moderate (15 minutes)
• Type	Extended production response
• Speededness	Speeded
Language of expected response:	
• Grammatical	General and technical vocabulary, varied grammatical structures
• Textual	Cohesive, well-organized piece of oral production

(continued)

Table 3 continued

• Functional	Ideational, manipulative (instrumental, interpersonal), heuristic
• Sociolinguistic characteristics	Standard dialect, relatively formal, natural language
• Topical characteristics	Topic selected by the examinee, has to be related to test-taker's area of specialization
Relationship between input and response	
• Reactivity	Reciprocal
• Scope of relationship	Broad
• Directness of relationship	Indirect

A performance test of oral ability with two parts – a 10-minute teaching part and a 5-minute question/answer part – was then developed. Both teaching and question/answer tasks closely simulated those of actual instructional settings and were thus believed to engage ITAs in performances from which their language abilities could be inferred. Like the test tasks, the rating instrument of the test (Appendix 2) was directly affected by the test constructs and included as many ability components as those in the construct definition. The performance of test-takers on each component was to be rated by a panel of 5 raters (2 ESL teachers and 3 undergraduate students) in terms of the levels of ability exhibited in performing the test task. A 5-point ability-based scale ranging from lowest ability level 'no production at all' to the highest level 'excellent performance' was developed for this purpose. Thus, the test kit included the rating instrument, the rating scales, a description of the ability components in the rating instrument, and the test itself (for details, see Saif, 2002: Appendices 2–5).

V Phase III: test consequences

In the third phase of the study, the test developed in Phase II was introduced into the actual teaching environment, i.e. a one-semester-long training program. A study at the institutional level was conducted to find out whether the test was having any beneficial washback effect on the areas predicted by the theoretical framework presented in this study and, if so, what form(s) washback would take.

Would it appear in the form of the material(s) reflecting different aspects of the test's objectives, efficient presentation of the course content by the teacher, a higher/lower achievement on the part of the learners, or a change of curriculum and/or policies in the university?

1 Method

All entering foreign graduate students ($N = 47$) participated in this phase of the study at the recommendation of FGS and their respective departments. Given the lack of evidence of their spoken language ability level, the old version of the SPEAK (Spoken Proficiency English Assessment Kit) test was administered to them about a month before the start of the program. The purpose of this step was to determine whether subjects had the general spoken language proficiency required for the TA program. Their performance on the SPEAK was tape-recorded and later rated by two trained ESL instructors who had gone through the step-by-step training process provided by the SPEAK kit. The passing score was 220 out of 300. In the next stage, about a week after the administration of the SPEAK, those ITAs with a score of 220 and over were required to take the new performance test. There were two considerations here: (1) to exclude from the experiment those candidates who already possessed the abilities measured by the new test, and (2) to obtain a set of scores for the candidates participating in the course for the purpose of comparison with their end-of-term scores on a parallel version of the test (in which candidates were required to do 10-minute teaching presentations on a topic different from what they had chosen the first time). The new test was administered by ELC over a period of one week and rated by 2 ESL teachers and 3 native-speaking undergraduate students from ITAs' respective departments. Altogether, 17 raters (15 undergraduate students from different departments, and 2 ESL teachers) were involved in the administration and scoring of the test.

As a result of the first administration of the test, 26 subjects were chosen to participate in a 12-week-long, 4-hours/week training program geared to the objectives of the test. The subjects were then randomly divided into experimental and control groups. While the control group continued with the regular orientation program, the experimental group underwent the one-semester-long language course. Ideally, the investigation of the test effects on ITAs' learning would have subjected the control group to the previous ITA course

offered by ELC, but this was not possible since at the time this study was underway, the only training program offered for ITAs was the TA orientation program offered by FGS. Still, it was worthwhile to retest the control group at the end of the course to see whether any change in the performance of the experimental group could be attributed to the instruction they had undergone, especially if the instruction proved to be test-related. The participants had been informed by FGS that a new test was being administered, but they were blind to the research questions.

The course was taught by one of the ESL teachers involved in the preliminary administration and scoring of the test. This same teacher had also taught the previous ITA course and was familiar with the speaking test used in the past. As discussed in Section II, this enhanced familiarity with the test and the context of assessment could result in what Bailey (1996) calls 'washback to the programme' by affecting the teacher's decisions with respect to the teaching content, teaching activities, and her methodology.

As for materials, a textbook (Smith *et al*., 1992) and supplementary videotape (Douglas and Myers, 1990) were chosen for the course. The book, which consisted of 10 units, focused on the most common teaching tasks in university classrooms. This particular textbook was chosen because it did not assume knowledge of a particular field of specialization and equally emphasized the use and practice of language skills, communication strategies, cultural sensitivity, and teaching skills. The videotape also focused on cultural aspects, communication strategies, and teaching skills. The new test, however, was primarily concerned with language abilities since, as determined by the survey in Phase I, the communication problems of ITAs were mainly language-related: according to the department stakeholders, compared with native-speaking TAs, ITAs were not at a disadvantage in the areas of topical knowledge and teaching abilities, and cultural differences were not highlighted by undergraduate respondents as a source of miscommunication in the classroom. Thus, the textbook that potentially reflected, promoted and practiced the areas of language ability measured by the test could also lead to an emphasis on topics irrelevant to the test if the teacher felt compelled to cover everything in the prescribed materials. It was, therefore, important to find out if the teacher was using the textbook and supplementary video in a way that promoted the abilities encouraged by the test. It was equally important to determine the teacher's approach to the course, her methodology, the type of in- and out-of-class activities and how they related to the test.

2 Results

To answer the above questions and to see whether, in general, the test was having the influence it was intended to have on the study's various participants (i.e. raters, teachers, ITAs), data were gathered at different times before, during, and after the training program using both quantitative and qualitative research methods. The data gathering methods as well as the results they produced are presented in the rest of this section.

a The raters: Following the first administration of the test, a preliminary survey was carried out among the raters. Information was gathered through a short questionnaire, observation of the testing sessions, and examination of the optional remarks made by ESL teachers during test administration. While the questionnaire was intended to elicit raters' views of the rating instrument and the rating process, the observation of the testing sessions focused on how, in practical terms, the raters reacted to the testing instrument and the rating process. It should be noted that the raters were not aware that their behavior was being observed by the researcher who, they believed, attended the testing sessions to ensure a problem-free videotaping of test-takers' performances. Table 4 summarizes the raters' responses to the questionnaire.

Observation results, in general, confirmed raters' responses to the questions above. All raters managed to score the test during or immediately after the examinee's presentation. Undergraduate raters occasionally verified the description of the rating categories; however, after the first few cases, all raters became more adept at the

Table 4 Raters' reaction to the performance test

	Yes	No
1) Do you understand all the ability components of the rating instrument?	76%	24%
2) Do you think the performance categories are adequate for measuring ITAs' spoken language ability?	59%	41%
3) Do you believe that the test is a practical one?	88%	12%
4) Do you believe that the 0-4 rating scale is reasonable and clear?	71%	29%
5) Do you regard the test task as closely related to the real-life tasks?	88%	12%
6) Do you believe that the test content would motivate ITAs to improve their spoken English?	94%	6%
7) Do you think that on-the-spot scoring is practical?	76%	24%

process and completed the rating instrument more quickly. Observations also revealed that the authenticity of the discussion topics together with the presence of the undergraduate raters familiar with the subject often led to a genuine discussion (especially during the question-answer part) which, in turn, enabled the panel to better evaluate the ability areas in question. ESL raters' written comments on individual examinees' performances further supported this point:

> [the test-taker] mostly read from the text, but then there was a dramatic change when answering the questions … textual knowledge and pronunciation need some work…
>
> … didn't quite answer the questions, … didn't understand what they were asking … several attempts, … definitely has comprehension problems.
>
> … wrote too much during the presentation, very little actual speaking until he had to answer the questions. (Saif, 2002: 156)

b The teacher: The reaction of the teacher to the test was examined through observations of class activities and a follow-up interview with the teacher. Observations, which covered the whole class period, were conducted in 6 sessions during the first, fifth, and eighth week of the course. Observation sheets – filled out by the observer – required the following information: reference to course objectives, materials used in class (i.e. the textbook, the supplementary video, any extra material), abilities emphasized and practiced in class, teaching activities, time allocated to different activities, nature and form of students' participation in class activities, and in-class evaluation. To avoid influencing the teacher's behavior in class, observations were carried out by a trained graduate student in applied linguistics; no recording instrument was used. Because of the long tradition of collaboration between the TESL program and ELC at UVic, ESL teachers at ELC are routinely approached by student-teachers who seek permission to observe ESL classes. It was thus quite normal for the teacher to have observers in her class every now and then. Subsequent analysis of observer's notes revealed that a number of aspects of the training could be traced to the test. These are summarized in Table 5 (direct references to the test are marked with an asterisk).

To find out why the teacher did the activities or made the choices recorded during the observations, a follow-up interview was held with the teacher. Since she was not aware of the specific research questions, in an unstructured interview she was asked to explain 'how' she taught the course and comment on course objectives,

Table 5 Summary of class observation results

	Week one		Week five		Week eight	
	1st period	2nd period	1st period	2nd period	1st period	2nd period
Course objectives	• unit 1: The section on course objectives replaced by a copy of the rating instrument distributed by the teacher* • teacher mentions that the final test would be a repeat of the initial performance test*	• functional language: • expressions used to open a discussion • expressions used to close a discussion	unit 4: teacher refers to Focus 1 section (content relevance) as the main objective of the unit in terms of one of the test's ability components (achieving communicative goal through production*	production and comprehension strategies to be used while teaching a group	• unit 7: presenting a topic • teacher presented the objective of this session as: strategies to use for long presentations*	session completely devoted to students' long presentations*
Materials	• textbook, unit 1: pronunciation and grammar sections • teacher emphasizes that course materials promote the skills listed on the rating instrument*	• unit 1: functional language section • section on cultural topics not discussed or referred to* • materials in CALL facility introduced for pronunciation practice	textbook, unit 4: sections on 'content relevance', 'methods of thought development', 'organizational clues' and 'cohesive devices'	• textbook, unit 4: 'content relevance' section • assignment: presentation by students • focus 2 (manner of speaking) assigned as out-of-class optional work*	• text book, unit 7 • assignment section on organizational skills	

(continued)

Table 5 continued

	Week one		Week five		Week eight	
	1st period	2nd period	1st period	2nd period	1st period	2nd period
Language abilities practiced	teacher defines the ability components on the rating instrument (as course objectives) and provides examples*	• functional language • pronunciation • grammar	setting communicative goals, and functional language: defining, explaining visuals, making analogies, comparing	• ability to communicate meaning in interactive contexts • vocabulary and grammar	• topical knowledge • pronunciation strategies • use of appropriate connectors	all abilities on the rating instrument
Teaching activities	• one-to-one tutorials recommended by the book replaced with group work and teacher/peer feedback* • group instruction: practical pronunciation tips, complex sentence structures (65 min)	• functional language: 'introducing yourself' • teacher's illustration: introduces herself 2 times (as an ESL teacher, and pretends to be an ITA in chemistry) • encourages ITAs to apply 1st period's pronunciation and grammar tips to their presentations (30 min)	• group instruction: teacher explains when and how to use different methods of thought development • provides numerous examples of conjunctions/ cohesive devices commonly used in each method. (80 min)	• teacher discusses common strategies to compensate for lexical and grammar problems • teacher and peer feedback forms in the textbook are not used for feedback • teacher distributes copies of the rating instrument for peer feedback (35 min)*	group instruction on strategies to: open a presentation (introducing the topic, its relevance to the course in general etc.), present the content (using notes, visuals, handouts), and conclude a discussion (by summarizing, drawing conclusion, making suggestions, etc.) (100 min)	

(continued)

Students' participation	• did the grammar exercise • asked questions about components of the rating instrument (e.g., register, topical knowledge) (20 min)*	students take turn to introduce themselves (60 minutes, max. 5 min each)	• did exercises on organizational cues, vocabulary and grammar • students told to choose a concept/term from their field for next sessions' presentations (20 min.)	in groups of three, each ITA tries to explain a technical concept using different strategies and methods of development. (60 min, max. 5 min each ITA)	• did the exercise on 'logical connectors' • contributed to teacher's discussion by offering ideas/examples (30 min)	• ITAs presented a 10-min talk on a topic they had prepared in advance. • teacher and peers filled out their feedback forms to be discussed the following session (100 min).
Evaluation	none	mock presentations	none	mock presentations	none	mock exam

Note: *Direct references to the test

materials, in-class teaching and testing activities, and ITAs' performances – the areas already addressed by the observations.

The teacher described her class activities as learner-centered and interactive in nature. She was convinced, based on ITAs' performance during the initial administration of the test, that ITAs had adequate knowledge of grammar and vocabulary; however, she thought they needed to improve their pronunciation, organization, and coherence. With respect to pronunciation and comprehension skills, she believed that given the time limit, the textbook strategies were very useful in facilitating presentation skills, yet she felt the need to supplement the text by directing the students to do out-of-class practice using resources that directly addressed such problems. At the same time, she had decided not to use the supplementary video in class. Instead, she had chosen to allocate the class time to what she called 'more didactic activities' such as group practicing of the communication strategies introduced by the book and providing feedback on the students' 'own' presentations. She explained she preferred group instruction because during the initial administration of the test she had noticed that the students had similar language problems. She, therefore, believed they could immensely benefit from the teacher's comments on their peers' performances. Also, the teacher was very satisfied with the students' participation in class activities especially during the instructor/peer feedback time. She described her role in such discussions as the provider of feedback – mostly in the form of strategies – to deal with the problems that interfered with the communication of meaning. She emphasized that while it was not possible to improve the students' pronunciation or reduce their accent considerably during a few months of training, it was possible to improve their language use skills by helping them use communicative strategies to make themselves understood. She thought mini-lectures and simulating teaching sessions had a significant effect on ITAs' progress and boosted their confidence as TAs. Nevertheless, when asked how she would teach ITAs had she not been bound by the exam and the prescribed material, the teacher expressed her preference for a smaller more relaxed seminar-type classroom mainly focusing on teaching skills.

c Learning outcomes: To further determine whether any learning had taken place, the test was administered to both experimental and control groups at the end of the training program. A quantitative study was performed using the subjects' scores on the 5 major

Table 6 Paired samples statistics for experimental and control groups

		Experimental (n = 13)		Control (n = 9)	
		Mean	s.d.	Mean	s.d.
Pair 1	GK1	2.3590	.4177	2.1000	.3536
	GK2	2.8744	.5701	2.1444	.3424
Pair 2	TK1	2.2846	.3579	1.9444	.4454
	TK2	2.7615	.4678	1.9333	.3571
Pair 3	FK1	2.4538	.3666	2.1222	.4236
	FK2	3.0154	.5460	2.1000	.3317
Pair 4	SK1	2.6154	.4902	2.4444	.4640
	SK2	3.0769	.5445	2.5556	.4640
Pair 5	SC1	2.1308	.5064	1.7444	.2964
	SC2	2.7138	.2774	1.7400	.2617
Pair 6	OP1	2.2615	.4032	1.9889	.3756
	OP2	2.8692	.5978	2.0556	.3678

ability areas – Grammatical Knowledge (GK), Textual Knowledge (TK), Functional Knowledge (FK), Sociolinguistic Knowledge (SK), and Strategic Competence (SC) – as well as the Overall Performance (OP) measured by the test. Table 6 displays the mean values and the standard deviations for the pairs of variables compared.

Subsequent results of the Paired Samples Test showed a significant difference between the mean values of the experimental group Time 1 and Time 2 performances for all ability areas (Table 7). The control group, on the other hand, did not show any significant progress in any one of the ability areas during that time frame (Table 8).

Table 7 Paired samples *t*-test for the experimental group

	Paired differences					t	df	Sig. 2 tail
				95% Confidence interval of the difference				
	Mean	s.d.	Std. error mean	Lower	Upper			
GK1-GK2	−.5154	.3349	.0929	−.7178	−.3130	−5.549.	12	.000
TK1-TK2	−.4769	.4070	.1129	−.7229	−.2310	−4.225	12	.001
FK1-FK2	−.5615	.4874	.1352	−.8561	−.2670	−4.154	12	.001
SK1-SK2	−.4615	.5189	.1439	−.7751	−.1480	−3.207	12	.008
SC1-SC2	−.5831	.3723	.1033	−.8080	−.3581	−5.647	12	.000
OP1-OP2	−.6077	.3685	.1022	−.8304	−.3850	−5.946	12	.000

Table 8 Paired samples *t*-test for the control group

	Paired differences					t	df	Sig. 2 tail
				95% Confidence interval of the difference				
	Mean	s.d.	Std. error mean	Lower	Upper			
GK1–GK2	−.0444	.0601	.0200	−.0906	.0017	−2.219	8	.057
TK1–TK2	.0111	.1537	.0512	−.1070	.1292	.217	8	.834
FK1–FK2	.0222	.1922	.0641	−.1255	.1700	.347	8	.738
SK1–SK2	−.1111	.3333	.1111	−.3673	.1451	−1.000	8	.347
SC1–SC2	.0044	.0947	.0316	−.0684	.0773	.141	8	.892
OP1–OP2	.0667	.1414	.0471	−.1754	.0420	−1.414	8	.195

Using the scores of the control and experimental groups in the Time 2 administration of the test, an Independent Samples *t*-test was then performed to test the equality of the means for the two groups of cases. The differences were found to be significant indicating that the experimental group not only showed progress relative to its own performance in the Time 1 administration of the test, but also performed significantly better than the control group in the Time 2 administration of the test. This was further supported by the results of the one-way ANOVA test and the GLM multivariate tests of between-subjects effects, both revealing a group effect on the performance of the learners in the second administration of the test (Table 9).

Table 9 Differences between groups in Time 2 administration

Variables	*T*-test for equality of means ($n = 22$)				GLM test of between-subjects effects (source: group)*	
	t	df	Sig. (2-tailed)	Mean difference	F	Sig.
GK2	−3.422	20	.003	−.7299	11.713	.003
TK2	−4.473	20	.000	−.8282	20.012	.000
FK2	−4.472	20	.000	−.9154	19.997	.000
SK2	−3.018	20	.004	−.5214	10.930	.004
SC2	−5.584	20	.000	−.9738	24.568	.000
OP2	−3.621	20	.002	−.8137	13.114	.002

Note: *The One-Way ANOVA test (Between Groups) produced identical results.

Table 10 Tests of within-subjects effects ($n = 22$)

Source	Measure	F	Sig.
Time	Grammatical knowledge	24.249	.000
	Textual knowledge	10.601	.004
	Functional knowledge	9.833	.005
	Sociolinguistic knowledge	8.467	.009
	Strategic competence	20.527	.000
	Overall performance	27.034	.000
Time group ($n = 22$)	Grammatical knowledge	17.160	.001
	Textual knowledge	11.637	.003
	Functional knowledge	11.520	.003
	Sociolinguistic knowledge	3.171	.090
	Strategic competence	21.162	.000
	Overall performance	17.401	.000

Multivariate tests of within-subjects effects also showed a significant time effect as well as a combined time and group effect on the performance of the learners in general (Table 10). The results were consistent with those of the ANOVA repeated measures procedure.

The significant time effect shows that there was a difference in learning taking place as a result of time. However, the results in Table 7 confirm that this was due to a significant change in the scores of the experimental group. The significant interactive effect indicates that there might be differences in learning for time and group combinations.

VI Conclusions

The results obtained from interviews, observations, and quantitative analysis of test scores suggested that the ITA test had some influence on classroom-related areas such as teaching content, teaching methodology, and students' learning. The results also revealed that the depth, extent and direction of the effect differed with the affected area.

The content of teaching seemed to be the area showing changes directly triggered by the test. This was in line with the results of previous studies on washback (see, for example, Wall and Alderson, 1993; Alderson and Hamp-Lyons, 1996; Cheng, 1997) that found the content of language teaching as the area readily susceptible to change as a result of tests. Class observations and teacher interview revealed that the teacher's adaptation for the ITA course of the materials available to her were based on two factors: the objectives of the course and her impression of the ITAs' language abilities after

the first administration of the test. Note that course objectives had been introduced with reference to the test objectives and the components of the rating instrument by the teacher on the first day of the class. The teacher's subsequent modification to the materials supported her emphasis on the test objectives: she did not go through the prescribed textbook chapter by chapter and paid less or no attention to the sections (like those discussing cultural topics) that did not practice the oral skills evaluated by the test. On the other hand, she routinely covered and expanded on selected exercises practicing common pronunciation problems, complex structures, organizational methods, and communication strategies. It is also important to note that the teacher's decision to modify the supplementary material was also based on the ITAs' performances on the first administration of the test during which the teacher had located certain areas of language problem. She replaced the supplementary video with students' in-class presentations accompanied by teacher and peer feedback, and introduced extra materials for both in- and out-of-class practice in listening comprehension and pronunciation. The test, therefore, appeared to have directly and extensively influenced the teacher's choice of both main and supplementary material used inside and outside the classroom.

Class observations and teacher interviews before and after the ITA course suggested that the teacher's methodology and the choice of class activities were, to a large part, adapted to the contents and goals of the test. The information gathered in Phase I showed no relationship whatsoever between the ITA course offered in the past and the content and purpose of the Speaking Test linked to it. This particular teacher commented that the Speaking Test was not valid for ITA purposes and that she had chosen to focus on the development of ITAs' 'teaching skills' in that course. Although the teacher's entire classroom behavior could not be unambiguously related to the test influence – due to the lack of observation data from her previous ITA teaching – evidence from observations supported by the teacher interview after the program revealed that the new ITA test had influenced her teaching in many respects. For example, the teacher's initial analysis of the ITAs' areas of strength and weakness was based on their performances on the test. This, in turn, affected her choice of 'group instruction' as the main teaching activity. She further customized the textbook according to the aims of the course (which themselves had been introduced with reference to the ability components in the rating instrument), introduced extra material, allocated a substantial amount of time to the students' presentations

and feedback sessions, and used the test's rating instrument for in-class evaluations and feedback. Considering that the teacher was personally in favor of a seminar-type course primarily concerned with teaching skills, these adaptations imply that the test content positively influenced some of the important decisions the teacher made with respect to her classroom activities.

These findings are somewhat different from those of previous studies (e.g. Wall and Alderson, 1993; Cheng, 1997) that found no straightforward connection between the test and how the teachers taught. It should be noted, however, that in the present study, only one teacher was involved in the teaching process. It is not clear whether most teachers working in the same environment would show a similar reaction to the test. Alderson and Hamp-Lyons (1996), for example, have found that TOEFL affects different teachers differently. Furthermore, in this particular context, the teacher's enhanced awareness of the test caused by her involvement in the test administration process, interaction with other raters, understanding of the rating process, and the ability components of the rating instrument were partially responsible for the changes she made to her teaching later during the program. This outcome appears to support Shohamy's (1992) argument that the involvement of 'the agents of change' (in this case, the teacher) in the test development process would promote the positive washback on the instruction.

The test's positive influence on teaching in this study would seem to have been responsible for the improved learning outcomes at the end of the course. The significant difference between the students' performance in the Time 1 and Time 2 administrations of the test reveals that learning had in fact taken place in all ability areas. The numbers do not draw a direct cause-effect relationship between the test and the ITAs' learning outcomes, but they are very suggestive. The superior performance of the experimental group compared to that of the control group in Time 2 administration of the test suggests that this change was the result of the instruction that took place in the meantime. Given that the objectives of teaching were geared to those of the test and that the teaching content and activities were in the direction of the test, it could be inferred that the test-related teaching in this study mediated improved learning outcomes. It should be noted that these results are based on data gathered at an institutional level. One should, therefore, be cautious about generalizing them to settings beyond the context of this study, although there is no reason to believe that the needs, objectives, and working environment of ITAs in other English-speaking universities are different from those of ITAs at UVic.

The above results showed that test effects on classroom teaching and learning can be – directly or indirectly – ascribed to a test whose development is informed by the needs of the learners and the objectives of the institution. However, the study also revealed that with a high-stakes test like the ITA test used here, so many factors originating from sources other than those in the classroom environment were critical for helping positive washback continue to happen once the test was in effect. Although all test users and stake-holders in the educational system acknowledged the significance of ITAs' language testing and training, the performance test has not yet been used on a regular campus-wide basis to screen international graduate students for TA jobs. This is because the introduction of the test seems to have rekindled some dormant non-test-related ethical, political, and financial issues that have to be resolved before the test can be used on a permanent basis. Among them are questions as to whether passing the test should be mandatory for the assignment of teaching assistantship to foreign graduate students; whether requiring prospective foreign graduate students to show proof of their spoken language ability would negatively affect the number of applicants to the university's graduate programs; or who should financially sponsor ITA language training, ITAs themselves, their departments, or FGS. So, while the test affected the educational system by bringing such issues to the attention of the administrators, any change in the educational system – as a result of the test – necessitates addressing these problems first.

On the whole, the results of the study indicate that while high stakes language tests that address the various needs of test takers and the educational system in general could positively affect teaching and learning activities, the test by itself cannot create change in the educational system. There exists an intricate web of different yet related factors that could enhance or interfere with a test's effects being realized as educational change. While the study reiterates the complexity of investigating washback noted by previous studies (e.g. Alderson and Wall, 1993; Andrews *et al.*, 2002), it also provides an indication as to the source of this complexity possibly being the concept of washback embracing both test effects and the educational changes resulting from such effects. However, the question is whether future research in washback – especially those studying deliberate attempts to create positive test washback – should resolve this complexity by making a distinction between the test 'effects' and the expected educational 'change' as a result of such effects. The former is a phenomenon that, based on the results of this study,

could be to a large extent predictable and attainable at the design level while the latter depends on various non-test-related factors in the educational system that are not always controllable by test-developers and users.

Acknowledgements

The theoretical model underlying this study was presented at the 2001 annual conference of the American Association of Applied Linguistics (AAAL), St. Louis.

I would like to acknowledge the doctoral fellowship from the Iranian Ministry of Culture and Higher Education, the financial and administrative support from the Faculty of Graduate Studies, UVic, and the grant from Grants and Awards Committee of the TOEFL Policy Council. I would also like to thank the three anonymous reviewers of *Language Testing* for their insightful comments and invaluable suggestions on an earlier version of this article.

References

Alderson, J.C. 1986: Innovations in language testing. In Portal, M., editor, *Innovations in language testing*. NFER Nelson, 93–105.

—— 1991: Language testing in 1990s: how far have we come? How much further have we to go? In Anivan, S., editor, *Current developments in language testing*. SEAMEO Regional Language Centre, 1–26.

Alderson, J.C. and **Hamp-Lyons, L.** 1996: TOEFL preparation courses: a study of washback. *Language Testing* 13, 280–97.

Alderson, J.C. and **Wall, D.** 1993: Does washback exist? *Applied Linguistics* 14, 41– 69.

Andrews, S., Fullilove, J. and **Wong, Y.** 2002: Targeting washback: a case-study. *System* 30, 207–23.

Bachman, L. and **Palmer, A.** 1996: *Language testing in practice: designing and developing useful language tests*. Oxford University Press.

Bailey, K. 1996: Working for washback: a review of the washback concept in language testing. *Language Testing* 13, 257–79.

Berwick, R. 1989: Needs assessment in language programming: from theory to practice. In Johnson, R.K., editor, *The second language curriculum*. Cambridge University Press, 48–62.

Brindley, G. 1989: The role of needs analysis in adult ESL program design. In Johnson, R.K., editor, *The second language curriculum*. Cambridge University Press, 48–62.

Cheng, L. 1997: How does washback influence teaching? Implications for Hong Kong. *Language and Education* 11, 38–54.

Douglas, D. and **Myers, C.** 1990: *Teaching assistant communication strategies*. Videotape and instructor's manual. Iowa State University Media Production Unit.

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

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

—— 1993: Backwash and TOEFL 2000. Unpublished manuscript. University of Reading.

Khaniya, T.R. 1990: The washback effect of a textbook-based test. *Edinburgh Working Papers in Applied Linguistics* 1, 48–58.

McNamara, T. 1996: Measuring second language performance. Longman.

Messick, S. 1996: Validity and washback in language testing. *Language Testing* 13, 241–56.

Morrow, K. 1986: The evaluation of tests of communicative performance. In Portal, M., editor, *Innovations in language testing*. NFER Nelson, 1–13.

Munby, J. 1981: *Communicative syllabus design: a sociolinguistic model for defining the content of purpose-specific language programmes*. Cambridge University Press.

Pearson, I. 1988: Tests as levers for change. In Chamberlain, D. and Baumgardner, R., editors, *ESP in the classroom: practice and evaluation*. ELT Document 128. Modern English Publications, 98–107.

Saif, S. 2002: A needs-based approach to the evaluation of the spoken language ability of international teaching assistants. *Canadian Journal of Applied Linguistics* 5, 145–67.

Shohamy, E. 1992: Beyond proficiency testing: a diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal* 76, 513–21.

Shohamy, E., Donitsa-Schmidt, S. and **Ferman, I.** 1996: Test impact revisited: washback effect over time. *Language Testing* 13, 298–317.

Smith, J., Meyers, C. and **Burkhalter, A.** 1992: *Communicate: strategies for international teaching assistants*. Regents / Prentice Hall.

Swain, M. 1985: Large-scale communicative testing. In Lee, Y.P., Fok, A.C.C.Y., Lord, R. and Low, G., editors, *New directions in language testing*. Pergamon, 35–46.

Turner, C.E. 2001: The need for impact studies of L2 performance testing and rating: identifying areas of potential consequences at all levels of the testing cycle. In Elder, C., Brown, A., Grove, E., Hill, K., Iwashita, N., Lumley, T., McNamara, T. and O'Loughlin, K., editors, *Experimenting with uncertainty: essays in honour of Alan Davies*. Cambridge University Press, 138–49.

Vernon, P.E. 1956: *The measurement of abilities*. University of London Press.

Wall, D. 2000: The impact of high-stakes testing on teaching and learning: can this be predicted or controlled? *System* 28, 499–509.

Wall, D. and **Alderson, J.C.** 1993: Examining washback: the Sri Lankan impact study. *Language Testing* 10, 41–69.

Watanabe, Y. 1996: Investigating washback in Japanese EFL classrooms: problems of methodology. In Wigglesworth, G. and Elder, C., editors, *Australian review of applied linguistics*. Series S 13. Applied Linguistics Association of Australia, 208–39.

Wiseman, S. 1961: The efficiency of examinations. In Wiseman, S., editor, *Examinations and English education*. Manchester University Press.

Appendix 1 ITAs' questionnaire

1. Participant's Identity:
   age (optional): sex: nationality:
2. Language
   first language: years of EFL/ESL training:
   TOEFL Score: TSE score (if available):
3. Purposive domain
   instructional/educational/research/other:
   discipline:
   years of teaching experience
4. Context of target language use
   classroom/laboratory/office/other:
   physical setting: University of Victoria, Department:
   frequency of target language use in this context:
   regularly/often/occasionally/seldom
   sociocultural characteristics of the setting
   professional/non-professional
   culturally familiar/unfamiliar
   formal/informal
   linguistically demanding/undemanding
   educationally developed/undeveloped
   other:
5. Interaction
   participant's position: graduate student/ITA
   people with whom the participant interacts:
   undergraduate students/peers/professors/other:
   number: individual/small group/large group
   age group: adult
   sex: male/female/mixed
   nationality: Canadian/other:
6. Type of communication
   spoken/written/production/comprehension/other:
   direction: bilateral/unilateral

Appendix 2 Rating instrument (*Source*: Saif, 2002)

Based on the test-taker's performance during phases 2 and 3, the raters will use the following rating instrument. Judgment may be based on the notes that the raters have taken during the presentation or by viewing the videotapes after the test is over. Raters should

review and completely understand the ability components listed here and the rating scale before administering the test.

Name: Date: Rater:

Directions: Please check only one number for each category.

Ability Levels	**None**	**Limited**	**Moderate**	**Extensive**	**Complete**
	0	**1**	**2**	**3**	**4**
Ability areas					
A. Grammatical knowledge					
1. Vocabulary					
2. Grammar					
3. Pronunciation					
B. Textual knowledge					
4. Cohesion					
5. Conversational organization					
C. Functional knowledge					
6. Use of ideational, manipulative and heuristic functions					
D. Sociolinguistic knowledge					
7. Dialect					
8. Register					
E. Strategic competence					
9. Goal-setting					
10. Use of verbal strategies					
11. Use of non-verbal strategies					
12. Achievement of communicative goal through production					
13. Achievement of communicative goal through comprehension					
F. Overall performance					

Source: Reprinted with permission from the Canadian Journal of Applied Linguistics